Generalizations to Corrections of Measurement Error Effects for Dynamic Treatment Regimes

by

(Zachary) Dylan Spicker

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2022

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Liqun Wang Professor, Dept. of Statistics, University of Manitoba
Supervisor(s):	Michael Wallace Associate Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
	Grace Yi Adjunct Professor, Dept. of Statistics and Actuarial Sciences, Dept. of Computer Science University of Western Ontario
Internal Members:	Audrey Béliveau Assistant Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
	Leilei Zeng Associate Professor, Dept. of Statistics and Actuarial Science, University of Waterloo
Internal-External Member:	Sharon Kirkpatrick Associate Professor, School of Public Health Sciences, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Dylan Spicker is the primary author for all of the work presented in this thesis. This work was conducted under the supervision of Michael Wallace and Grace Yi.

The work in this thesis consists, in part, of material from one published manuscript (Chapter 6), and two as-yet unpublished manuscripts (Chapters 3 and 4). The presented work in this thesis expands beyond that included in the manuscripts, both expositionally and substantively. This was authored solely by Dylan Spicker, with structural and phrasing assistance from their supervisors.

Abstract

Measurement error is a pervasive issue in questions of estimation and inference. Generally, any data which are measured with error will render the results of an analysis which ignores this error unreliable. This is a particular concern in health research, where many quantities of interest are typically subject to measurement error. One particular field of health research, precision medicine, has not yet seen a substantive attempt to account for measurement error. Dynamic treatment regimes (DTRs), which can be used to represent sequences of treatment decisions in a medical setting, have historically been analyzed assuming, implicitly, that all quantities are perfectly observable.

We consider the problem of optimal DTR estimation where quantities of interest may be subject to measurement error. The nature of this problem is such that many existing techniques to account for the effects of measurement error need to be expanded in order to accommodate the data which are available in practice. This expansion further highlights theoretical shortcomings in the existing methodologies.

This thesis begins by expanding existing methods for correcting for the effects of measurement error to accommodate issues which are frequently observed in real-world data. We expand the most commonly applied measurement error corrections (regression calibration and simulation extrapolation), demonstrating how they are able to be conducted with non-identically distributed replicate measurements. We further expand simulation extrapolation, which typically assumes normality of the underlying error terms, proposing a nonparametric simulation extrapolation. These expansions are conducted generally, separate from the specific context of optimal DTR estimation.

Following the expansion of these extant techniques, we consider the problem of errors in covariates within the DTR framework. We apply the aforementioned generalized error correction techniques to this setting, and demonstrate how valid estimation and inference can proceed. Finally, we consider problems which are present when there is treatment misclassification in DTRs, proposing techniques to restore consistency and perform valid inference. To our knowledge this work represents the first substantive attempt to explore these problems. Thus, in addition to proposing methodological solutions, we also elucidate the particular challenges of estimation in this setting. All proposed techniques are explored theoretically, using simulation studies, and through real-world data analyses.

Acknowledgements

I would like to thank my supervisors, Michael Wallace and Grace Yi, for their help, mentorship, insight, and encouragement during my studies. I feel incredibly lucky to have gotten to start my journey as a researcher under your supervision, and will always hold close the lessons you have taught me.

I would like to express gratitude to Drs. Audrey Béliveau, Sharon Kirkpatrick, Liqun Wang, and Leilei Zeng for their service on my examining committee.

My work would not have been possible without the support of my teachers, professors, and mentors who have guided and encouraged me along the way. I am indebted to my parents, siblings, and other family who have provided inspiration, motivation, and a tremendous amount of patience during (frequent, likely incomprehensible) discussions of my work. I want to thank my friends and peers at Waterloo for the wonderful support network, which has made these efforts exciting and fun. It has been an absolute pleasure to work alongside you all. Kelly, thank you for helping to unlock the secrets to research with me (the answer is always Taylor series, of course). For my friends outside of academia, thank you for the escapes that you provide, the wonderful conversation, and your continual understanding. You all give me constant reminders of what truly matters. Celeste and Dana (and Earl!), thank you for being the best neighbours and friends that anyone could ask for.

Finally, I need to thank my partner, Melissa. Without your love and support none of this would have been possible. I cannot overstate my appreciation for you. My successes are truly yours. Thank you.

I further want to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding throughout my doctoral research.

Dedication

For Charles and Sadie...
(... and Garth, wherever you are, I hope you're partying on).

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Measurement Error and Precision Medicine

In this thesis we will address problems related to measurement error and precision medicine. We begin by introducing both measurement error and precision medicine at a high-level, before considering the specific problems that this thesis will address. Measurement error refers to any scenario where a variable of interest cannot be accurately observed. If the discrepancy between the truth and what is available is stochastic, we will refer to this as measurement error. Our definition excludes systematic errors, where measurements deviate in a deterministic manner. There are many possible causes of measurement error. For instance, the instrument that we are using to take measurements may have random error inherent in it (as is the case with using calorimetry for evaluating the mass of nuclear materials [83]), or it may be an innately immeasurable quantity (as is the case when we are interested in someone's underlying blood pressure, and can only take discrete measurements at clinical visits [88]).

While it is generally best practice to minimize the occurrence of measurement error, it is not typically possible to eliminate errors entirely. This matters because of the "Triple Whammy of Measurement Error", [1] which specifies that measurement error in covariates:

1. Causes bias in parameter estimation.

[1] A phrase coined in Carroll, Ruppert, Stefanski, and Crainiceanu [7, p. 1]; they originally referred to the first two components as the "Double Whammy of Measurement Error", but emphasize the masking effect in the second version of their book.

2. Leads to a loss of power.

3. Masks possibly interesting features of the data.

There have been many methods proposed to correct for these effects of measurement error. The goal of a measurement error correction is to restore analytical guarantees to the underlying analysis, so that any estimation or inference can be relied upon. Providing a comprehensive list of correction techniques is infeasible (see, for example, Carroll, Ruppert, Stefanski, and Crainiceanu [7], Yi [101], Buonaccorsi [6], and the included references). This thesis will focus primarily on two widely used techniques: regression calibration [10, 31] and simulation extrapolation (SIMEX) [16]. We also consider how likelihood and estimating equations are leveraged to correct for the effects of measurement error.

As a clarification on language we often speak of "measurement error correction techniques." This is a convenient shorthand to describe techniques which are designed to reduce, or eliminate, the impacts of measurement error in an analysis. It is more accurate to describe such techniques as "corrections for the impacts of measurement error", though the shortened "measurement error corrections" can be a convenient stand-in. It is important to note that, despite the occasional use of this language, these techniques are not eliminating or addressing the errors directly: once measurement error is present, it remains so. Instead, these are techniques that are designed to account and adjust for the induced bias or loss of power that would otherwise be present if the errors were ignored.

Measurement error correction techniques have been used frequently in health and medical statistics. The nature of health data is such that many quantities of interest (including blood pressure [88], dietary intake [89], or smoking status [60]) are commonly subject to error. As such, many of the recently developed measurement error corrections have been motivated through a biostatistical lens. In parallel with these developments, the field of *precision medicine* has emerged as central to evidence-based medical practice.

Precision medicine is the practice of using patient-specific information (such as demographic, genetic, or lifestyle factors) to tailor the treatment that a patient receives [49]. This approach stands in contrast with a disease-centric view of medicine, where the best treatment for a particular condition is sought. Generally, medical researchers are interested in determining causal relationships between treatments and outcomes. When these relationships are allowed to be mediated by patient characteristics, this is causal inference for the purpose of precision medicine.

One framework for formalizing causal inference in this setting is through the use of dynamic treatment regimes (DTRs). Broadly, a DTR is a (set of) decision rule(s) which take as input patient information and produce as output a treatment decision. These stand in

contrast to static treatment regimes, which do not tailor their outputs based on the particular patient. DTRs are capable of encompassing both acute, *single-stage* treatments (for instance, deciding on the best treatment during a clinic visit where the patient is exhibiting flu-like symptoms), and longitudinal, *multi-stage* treatments (for instance, providing ongoing therapy for a patient with depression). When considering DTRs, interest may be in the estimation of the optimal treatment rules, in the prediction of the optimal outcome, or in the assessment and comparison of specific treatment strategies (relevant methods for DTRs are summarized in Tsiatis, Davidian, Holloway, and Laber [90] and Chakraborty and Moodie [14]).

Despite the parallel development of methods for correcting for the effects of measurement error, and methods for conducting precision medicine, there has not been a substantial attempt to integrate these two fields of study. When we consider the known impacts of measurement error generally, and the degree to which measurement error is a problem in health research, it is sensible to assume that many of the same issues will need to be addressed in the context of precision medicine. Addressing this gap is the motivating goal for the work conducted throughout this thesis.

1.2 Problems of Interest and Thesis Structure

The problem of assessing, and correcting for, the impacts of measurement error in dynamic treatment regimes is a problem with several important branches. It is important to consider where the errors are: the impacts and required corrections are different if errors are in the tailoring variables as compared to the treatment indicators. There has been a growing field of literature considering measurement error in causal inference generally. While some of this research is applicable to our setting, the added complexity of DTRs necessitates further consideration.

The estimation of optimal DTRs is conceptually similar to standard regression procedures. Despite this underlying simplicity, much of the foundational literature establishing methods for optimal DTR estimation is presented for methodological researchers, relying on deep, mathematical sophistication, rather than for prospective practitioners [59, 70]. Over time there has been a more concerted effort to increase the accessibility of these techniques, both through more accessible communication (as in the monographs by Chakraborty and Moodie [14] and Tsiatis, Davidian, Holloway, and Laber [90]) as well as through the development of methods that are more explicitly grounded in familiar techniques (as with Wallace and Moodie [95]). The observed "research-practice gap" [90] within the DTR lit-

erature emphasizes the need for techniques which are accessible to applied researchers for the continued success of the field.

Within the context of measurement error corrections broadly, this need has been further demonstrated in a recent study. In an analysis of applied literature, Shaw et al. [80] concluded that, at least in the fields that were considered, when measurement error corrections were used, they tended to be corrections that are comparatively straightforward to implement. Most analyses that they looked at made no meaningful correction for the impacts of measurement error, and when researchers did, they tended to leverage techniques which are less mathematically principled but far more accessible. These findings, alongside the aforementioned research-practice gap, underscore the importance of considering accessibility at the stage of methodological development.

Against this backdrop, this thesis will balance two goals in the pursuit of understanding and addressing the impacts of measurement error in the estimation of optimal DTRs. On one hand, we wish to rigorously approach these problems and provide principled guidance which is theoretically grounded. On the other hand, we wish to approach these problems using methods which are comparatively straightforward, with an eye towards those which can be made to be accessible. This desire for "rigorous accessibility" will be extended beyond the context of measurement error in DTRs, as we try to achieve this balance within the context of general measurement error correction procedures. While the focus will be on ensuring that the techniques are comparatively straightforward to implement, at least for those who are generally familiar with the subject matter that we cover, the primary aim of this thesis is the mathematical justifications for these accessible techniques. As such, much of what follows will focus on communicating the mathematical ideas, rather than serving as a tutorial for the implementation.

This thesis is structured primarily around four different questions.

1. In Chapter 3 we introduce methods for generalizing some commonly applied measurement error correction techniques, so that they are useful in a wider range of settings. These generalizations were initially motivated through shortcomings with the existing techniques, when attempting to apply them to data used to estimate optimal DTRs. We present the results generally here in service of the goal of accessible, defensible methods for measurement error correction.

2. In Chapter 4 we develop methods for extending a commonly applied measurement error correction technique to the case where the measurement error is not assumed to be normally distributed. Just as with the previous point, these results are motivated via problems observed in actual data. They are presented generally, as a way of broadening the settings in which these techniques can be applied with principle.

3. In Chapter 6 we investigate the impacts of, and propose corrections for, measurement error in the estimation of an optimal dynamic treatment regime. Specifically, this chapter considers the problem of *errors-in-variables*, for the factors that we wish to use as *tailoring variables* to cater treatment recommendations with. In this chapter we propose a correction procedure that permits the valid estimation of an optimal DTR when tailoring covariates are subject to measurement error.

4. In Chapter 7 we consider measurement error in dynamic treatment regimes, where the errors are in the assigned treatments. We consider this as a problem of *nonadherence*, where an individual may not (correctly) take their prescribed treatment and so there is a disconnect between the recorded treatment in the available data, and the true treatment that the individual took. In this chapter we explore the impacts of this setting, and describe a method for correcting for the bias that this form of measurement error introduces.

Within each of these chapters we first motivate and contextualize the problem, before briefly presenting the proposed techniques in a summarized manner, indicating the key implementation details. Following this, we explore in depth the theoretical concerns related to each problem, and demonstrate the utility of the proposed methods through simulations, and applications to real-world data analyses.

The remainder of the thesis is structured as follows. We first finish the introduction, presenting four motivating examples: the Framingham Heart Study (FHS) [44], The Korean Longitudinal Study of Aging (KLoSA) [99], the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) [26, 76], and the Multicenter AIDS Cohort Study (MACS) [45]. These four studies represent the four data analyses used within each of the aforementioned chapters, respectively. We introduce them as a means of motivating the problems we will consider, and provide more details within each chapter where they are analyzed. In Chapter 2 we discuss the methodological background required for measurement error corrections. This includes a discussion of measurement error models, regression calibration, SIMEX, and estimating equation correction techniques. With the background established, Part I is completed with Chapters 3 and 4. Part II explores dynamic treatment regimes, starting with Chapter 5 which introduces the background methodology for DTR estimation. This includes a discussion of potential outcomes, the formal DTR framework, and optimal DTR estimation through Q-learning, dynamic weighted ordinary least squares (dWOLS), and G-estimation. Following this, Chapters 6 and 7 are presented, building on the introduced concepts. Finally, Chapter 8 contains a brief discussion and summary of the presented work. Additionally, we include several appendices. In Appendix A, we present

background theory on M-estimation, a technique used widely throughout this thesis. Appendix B gives proofs for all of the theoretical results throughout. Appendix C provides a worked example of non-regularity in dynamic treatment regimes, a technical concern that we address with our methods but which is otherwise mathematically involved. Appendix D provides additional simulation results for the work conducted in Chapter 6.

1.3 Motivating Examples

1.3.1 Framingham Heart Study

The Framingham Heart Study is a large cohort study, which investigates the development of coronary heart disease (CHD) [44]. In the measurement error context, systolic blood pressure (SBP) is typically treated as an error-prone variable, and interest concerns the impact of the long-term average systolic blood pressure on the development of CHD [7].

The subset of the data that we will analyze follows 2876 individuals, aged 32–69, across three separate examinations. We take the patients' sex, age, and smoking status to be error-free, and assume that the serum cholesterol levels and systolic blood pressure are prone to error. These data are subject to incomplete replication. Of the 2876 total participants, systolic blood pressure measurements were available for all patients at the first visit, but missing for 153, and 390 patients at visits two, and three respectively. For cholesterol, at visits one, two, and three, there are 26, 256, and 538 patients without replicate measurements respectively.

Measurement error techniques are required to accurately analyze the FHS. This is because research interest is in the impact of long-term average SBP and serum cholesterol on CHD. However, long-term averages are typically immeasurable quantities. The measurements which are taken at clinical visits will be subject to measurement error. Blood pressure, for instance, is known to have both daily, as well as seasonal variation, distorting the measurements at any one clinic visit [7]. While previous analyses that we follow typically assume that each measurement of the underlying quantities corresponds to a repeated measurement, subject to the same error process, these data demonstrate that this is not the case. As a result, this study provides a comparatively simple example for motivating the need for relaxations to the assumptions made in replicate-based measurement error corrections. Specifically, we will use these data to motivate measurement error corrections that function based on any set of proxy measurements of the truth, rather than relying on these proxies to follow the same distributions.

1.3.2 Korean Longitudinal Study of Ageing

The Korean Longitudinal Study of Ageing is a longitudinal survey which was started in 2006, and follows South Korean citizens who are aged 45 and older. The survey is conducted by the Korea Employment Information Service, at two-year intervals, and seeks to determine the health effects of aging. Complete details regarding the survey are available on the study's website: https://survey.keis.or.kr/eng/klosa/klosa01.jsp.

The sample of the KLoSA that we consider follows from the analysis of Xu, Kim, and Li [99]. We consider a sample of 9842 individuals, with an interest in determining whether body mass index (BMI)[2] is predictive of hypertension in this population. However, within the KLoSA most individuals self-report their body weight and height, which is then used to compute their BMI. For a sub-sample of 505 individuals we have clinical measurements of these quantities in addition to their self-reported values. From this we can see that, not only is there substantial error present in self-reported information, but also that these errors are far from normally distributed. Many commonly applied correction techniques assume that errors are normally distributed; the KLoSA makes clear that this is not always a defensible assumption.

As a result, in conducting this analysis, measurement error techniques which can use validation data, and which do not assume normality of the errors are required. We use this setting to motivate a nonparametric version of a commonly applied measurement error correction technique, simulation extrapolation.

1.3.3 Sequenced Treatment Alternatives to Relieve Depression

The STAR*D study was a multistage randomized controlled trial, comparing different treatment regimes for patients with major depressive disorder. [26, 76] The study was split into four levels (with level two further subdivided into two sublevels) where, at each level, different treatment options were available to patients based on preference and progression through the study. At level 1, all patients were prescribed citalopram. The patients who entered the second level had seven treatment options available, characterized by 'switching' from citalopram to one of four other treatments options, or 'augmenting' treatment by receiving citalopram alongside one of three new treatments. The patients who continued to progress into levels three and four were offered similar treatment options. This is summarized in Figure 1.1.

[2] An individual's BMI is given by their weight (in kilograms) divided by their height (in metres), squared.

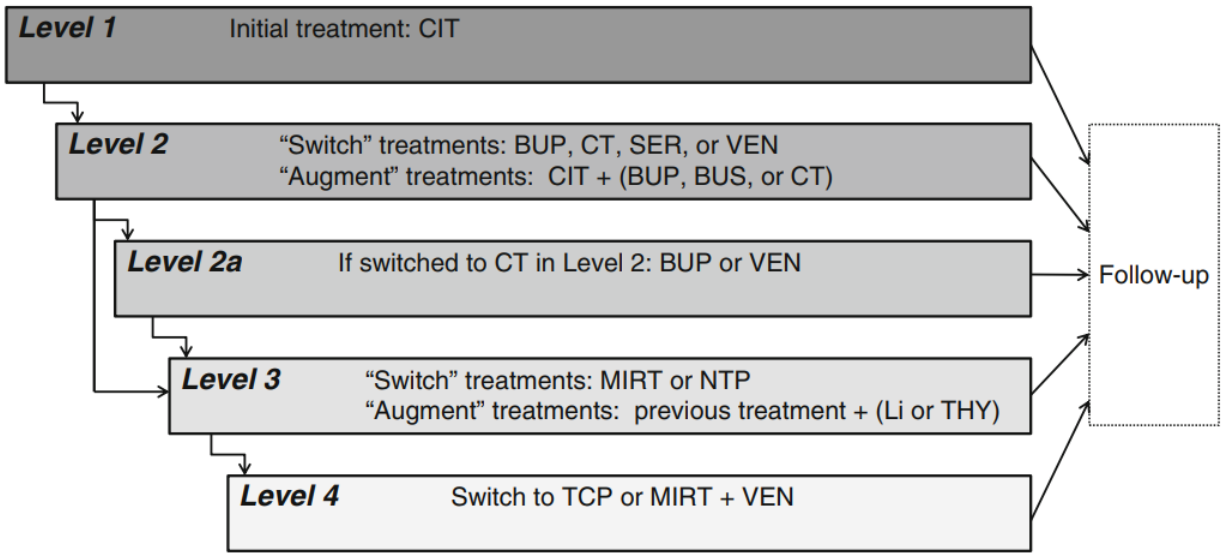

Figure 1.1: A flowchart (from Chakraborty and Moodie [14]), demonstrating the possible progress for a patient through the four phases of the STAR*D trial.

The severity of depression was measured through the Quick Inventory of Depressive Symptomatology (QIDS) score, where assessment was conducted during each phase both by the patient (denoted QIDS-S) and by a clinician (denoted QIDS-C). At the end of each study phase, patients who had a clinician assessed QIDS score less than or equal to 5 were considered to have entered remission, and were subsequently removed from the study. Only those patients who had not entered remission were moved to further levels.

When analyzing the data, generally the focus is on levels 2 and 3 (for instance in Chakraborty, Laber, and Zhao [13]). The treatment options are simplified into those which contain a selective serotonin reuptake inhibitor (SSRI) and those which do not. The goal of the study is to determine the best treatment options to reduce a patient's QIDS score, where their treatment preferences and starting symptomatology are taken into account. There are 283 patients in the sample who have all necessary measurements taken.

STAR*D motivates the need for measurement error techniques in the analysis of DTRs. Truly assessing the severity of depression in a patient is a task which, intuitively, may be subject to error. Since both the patient and the clinician take QIDS measurements, we can (and do) see that there is a disparity between these measurements in general, meaning that at a minimum, one of the two is error-prone. Despite this, errors have traditionally been ignored in the analysis of these data, motivating the questions we are seeking to answer.

1.3.4 Multicenter AIDS Cohort Study

The MACS study was a longitudinal cohort study investigating the impact of the HIV-1 infection in gay and bisexual men, which ran from 1984 to 2019 [45].[3] The study contained information on over 7000 men, and collected biological and behavioural data from participants every six months. The information collected ranged from demographic and psychosocial characteristics, through to detailed lab reports on blood samples from the individuals. While not specifically designed for the purpose of estimating a dynamic treatment regime, these data can be viewed as an observational study for assessing treatment options for patients with HIV/AIDS.

We consider an analysis of a subset of these data which are publicly accessible, and which look at a particular antiretroviral treatment, Zidovudine (AZT). The subset we use consists of information from 2929 patients, representing a total of 9316 visits. We consider a two-stage DTR regarding the timing of AZT prescription. We want to assess whether a particular patient should be started on an AZT regimen, at a given clinical visit, based on relevant demographic characteristics, as well as on the results from their lab reports. The primary outcome of interest is the CD4 cell count. CD4 cells are important cells for measuring the severity of an HIV infection as the virus attacks and destroys the cells. The fewer cells present in a patient's blood, the more severe the infection has grown. Our analysis primarily draws on that of Wallace, Moodie, and Stephens [96].

Starting in 1996, the MACS study began to collect information regarding patient adherence to treatment. These data demonstrate that, for a subset of the patients in our sample, we can see that adherence to prescribed AZT therapy is not perfect. As a result, an analysis which uses the prescription of AZT as the relevant factor will be subject to bias induced from the error in this variable (when used as a proxy for treatment itself). Nonadherence like this is a common problem across many studies, particularly when they are observational, and treatment is self-administered. Interestingly, the nonadherence present in the study is minor, and so an analysis of the MACS study serves as an indication of the importance of developing methods to account for the impacts of nonadherence.

[3] As of 2019, MACS and the Women's Interagency HIV Study (WIHS) combined to form the MWCCS.

Part I

Measurement Error

Chapter 2

Methodological Background: Measurement Error

In this chapter we formally introduce measurement error. We begin by discussing *measurement error models*, which capture the set of assumptions we make about the mathematical properties of error. We then introduce the data requirements for addressing measurement error. We discuss in detail both regression calibration and SIMEX, and introduce techniques for correcting for the impacts of measurement error based on unbiased estimating equations. We then discuss *misclassification*, which generalizes the concept of measurement error from a continuous setting to a discrete (or categorical) setting.

For the context of measurement error, we consider interest in a random variable X, which is not directly observable. In place of X, we observe X^*, which is called the surrogate (or observed version) of X. We also describe X^* as a proxy measurement. We say that X is *error-prone*. When interest lies in a numeric outcome (such as is the case with regression models), we take the outcome to be the random variable Y, and assume that it is measured without error. All other variates of interest, which are not subject to error, are denoted Z. We use U to represent the error process. That means that U is generally unobservable, and we are only interested in U to the extent that it distorts our measures X^* of X. For ease of exposition, we use notation which implies scalar values for these quantities, though most of the presented material is applicable to vector-valued variables.

When considering misclassification, we remain interested in a random quantity X which takes on discrete or categorical values. Generally, X is unobservable and in place we observe X^*, where X^* may not equal X. Because X and X^* are assumed to be discrete, we do not conceptualize the error process as U, but rather consider a probability distribution

that dictates how the misclassification process occurs. Again, we are interested in the misclassified measurements only to inform us of the true, underlying relationships. The selected notation is (once again) scalar-valued, but can be readily expanded to vector-valued random quantities as well.

2.1 Measurement Error Models

Typically the source of measurement error is considered irrelevant, except insofar as the mechanism may inform what mathematical models are justifiable. We tend to focus on classifying measurement error based on how it will be modelled, the so-called *measurement error model*. Generally, a measurement error model refers to the collection of assumptions that define how measurement error impacts our observations. While it is possible to conceive of many such assumptions, in practice several measurement error models are typically considered. These can be broken down based on structural and independence assumptions. Briefly, we will make assumptions regarding:

1. Whether we wish to treat the true values, X, as random quantities (called *structural modelling*), or as fixed parameters or random quantities with unspecified distributions (called *functional modelling*).

2. Whether we model the conditional distribution of X^* given X (called *classical error models*) or X given X^* (called *Berkson error models*).

3. The structural relationship between X and X^*. For instance, we could assume that $X^* = X+U$ (an *additive model*), or we could take $X^* = XU$ (a *multiplicative model*).

4. Whether we assume that our error process and outcome are independent (called *non-differential error*) or not (called *differential error*), given the true covariates.

5. The distributional and moment properties of U, for instance, we may assume that $E[U|X] = 0$ (for an additive model), $E[U|X] = 1$ (for a multiplicative model), or $\text{var}(U) < \infty$. Some methods may impose normality on U.

The set of assumptions that are made with respect to these properties will be referred to, collectively, as a measurement error model. The literature tends to focus on non-differential, classical, additive models, with or without distributional assumptions on U.

Before discussing the specifics of these assumptions, we need to differentiate between two philosophies: structural and functional modelling. Our introduction lends itself to

the structural approach, where the true values are simply seen as unobserved random quantities. In the functional approach, all inference is conditional on X. In this sense we can think of the unobserved true values as parameters instead of random variables. Certain correction techniques require taking one approach or the other, and certain techniques can be framed in both. Generally speaking, structural approaches have the drawback of requiring (nearly) correct distributional modelling, but provide more efficient estimators when this is possible. The functional approach is robust to these assumptions, but may be less efficient [101]. We will take the structural approach for exposition. That is, we will assume that the underlying, true quantity X is a random quantity, and we wish to make inference regarding its distribution.

Given a structural framing, it is often natural to think of X^* as a version of X that has been perturbed by U. Consider the measurement of blood pressure, where it seems likely that our measurement is the combination of the true blood pressure and the noise process. In other settings, it may be more reasonable to assume that X is a version of X^*, perturbed by U. For instance, if we are concerned with the impact of herbicides on plant growth, then we may expose plants to a known quantity of herbicide [75]. The dose of herbicide that ends up on each plant will likely not be precisely what we intended, leading to X being a perturbed version of X^*. The former is classical error, and the latter Berkson error. Our focus will remain entirely on classical error.

In terms of structural assumptions relating X and X^*, we could conceive of any function, say $X^* = g(X, U)$. In practice, most of the measurement error literature assumes that $X^* = g(X, U) = X + U$. While presenting existing methods, we will make this assumption of additive error. However, one key problem we will address in this thesis concerns broadening the class of structural assumptions that common methods can accommodate.

When interest lies in an outcome Y, error correction techniques tend to require the assumption of non-differential error (that is, Y and U are taken to be conditionally independent, given X). This may not be reasonable if, for instance, Y is measured prior to X^*, such as may be the case where X^* is self-reported smoking behaviour and Y is a lung cancer indicator [7]. This thesis will assume non-differential error. Additionally, we often want our measurements to be *unbiased* in the sense that $E[X^*] = E[X]$, or that $E[X^*|X] = X$. This amounts to assuming that $E[U|X] = 0$ in the additive case, and that $E[U|X] = 1$ in the multiplicative case. We often strengthen this and assume that U and X are independent, with $E[U] = 0$ (in the additive case). There may be situations where taking a strict distributional assumption (typically U to be normal) is required.

2.2 Data Requirements for Error Effects Correction

In order to correct for the effects of measurement error we require additional information to determine the size and structure of the error [7]. It may be the case that there is existing knowledge of the error distribution, such as having an estimate for $\text{var}(U)$ based on past research. However, we usually rely on the presence of additional data which we can exploit to estimate the required parameters. There are four main types of auxiliary data:

1. **Internal Validation data:** For some subset of the sample we observe the complete set of (Y, X, X^*). This allows for us to model the relationship between X and X^* explicitly, and correct for the observations without a corresponding X value.

2. **External Validation data:** In addition to our sample observing (Y, X^*), we have an external sample where we have observations for (X, X^*). This allows us to model the relationship between X and X^* explicitly in the external sample, and transport this to the main sample.

3. **Replicate data:** For some subset of the sample we observe multiple values of X^*. This allows for us to use decomposition of variance techniques to determine the size and structure of the error.

4. **Instrumental data:** For some subset of the sample we observe an additional covariate T, which is related to both Y and X^*, in an exploitable way (see below).

Validation data are ideal, but are relatively uncommon in practice. An external validation sample is useful only when we are willing to make a *transportability assumption*, taking the measurement error mechanisms to be the same in the main study and previous validation sample. Replicate data are more commonly found, and are particularly useful when assuming classical error models.

A variable T is called an instrumental variable if (1) it is **not** independent of X, (2) it is uncorrelated with the error U, and (3) it is uncorrelated with the residual error in Y once accounting for X (that is, T and $Y - E[Y|X]$ are uncorrelated) [7]. In this case, T is related to Y and X^* only through X. This allows for modelling of T given X^*, which can in turn be used to correct for the effects of error [7, 30, 86]. It is not always easy to verify whether the assumptions of an instrument are met, and assuming a variable is instrumental erroneously can cause significant issues in an analysis [7].

When auxiliary data are not available, and when there is no knowledge of the error distributions, it may still be possible to estimate the impact of measurement error through

sensitivity analyses [101]. In this case, we take a number of plausible measurement error models and the parameters which specify them, and correct our analysis assuming that they are the true measurement error models. From here we can see how sensitive the results are to varying degrees of measurement error, allowing some quantification of the uncertainty in our analysis. We will make different assumptions regarding the availability of auxiliary data, depending on the specific correction we are discussing.

2.3 Correcting for the Effects of Measurement Error

It is common to categorize the many existing correction methods on the basis of underlying methodology. For instance Fuller [30] summarizes corrections for linear models, and Carroll, Ruppert, Stefanski, and Crainiceanu [7] work with nonlinear models. For the purpose of this thesis, a separate distinction will be of particular importance: whether the correction aims to be consistent or *approximately consistent*. Consistency is often achieved in the measurement error literature through either complex calculations or strong assumptions. Particularly with our focus on precision medicine, there are circumstances where easy to implement methods which provide some protection against the negative impacts of measurement error are preferable to exact methods which are cumbersome to use. This is especially the case when methods cannot be implemented in standard software. For this thesis, much of our focus will centre on the approximately consistent methods of regression calibration [10, 31] and simulation extrapolation [16]. We will also consider exactly consistent techniques based on unbiased estimating equations [61, 101, 7].

2.3.1 Regression Calibration

Regression calibration[1][10, 31] functions by posing a model for X, given X^*, and then using this model to impute values of X, which can be used in a standard analysis. If a validation sample were available, then this can be achieved through direct modelling. Here, we would directly model X from X^*, filling in the records which do not have an observation. In this setting regression calibration can be viewed as an unsophisticated imputation technique since X can be considered missing [7]. If instrumental data are available, and the instruments are unbiased, then it may be the case that regressing T on

[1]In the measurement error literature, the phrase "regression calibration" has been used in a variety of different ways, applying to methods which are more or less related. Our terminology is taken from more recent monographs [7, 6, 101].

(X^*, Z) is the same as regressing X on (X^*, Z). This presents another direct modelling problem.

When there is no validation sample taken, and when unbiased instruments are not readily available, it may still be possible to model $E[X|X^*, Z]$. One common technique for doing this, when replicate measurements are assumed to be available, is through the use of the *best linear unbiased prediction* (BLUP). The BLUP approximates $E[X|X^*, Z]$ as the unbiased, linear function which minimizes the mean squared error (MSE). If we consider linear estimators of X, we take $\widehat{X} = \mu + \beta X^* + \gamma Z$, and select (μ, β, γ) such that $E[(X - \widehat{X})^2]$ is minimized. Taking μ_A to be $E[A]$, and Σ_{AB} as the covariance between A and B, the optimal parameters solve

$$\begin{bmatrix} \mu_X - \mu - \beta\mu_{X^*} - \gamma\mu_Z \\ \Sigma_{XX^*} - \beta\Sigma_{X^*X^*} - \gamma\Sigma_{ZX^*} \\ \Sigma_{XZ} - \beta\Sigma_{X^*Z} - \gamma\Sigma_{ZZ} \end{bmatrix} = 0.$$

If we assume classical, additive measurement error, such that

$$X^* = X + U, \tag{2.3.1}$$

with $E[U|X] = 0$, then some of these terms can be simplified. In particular, $\mu_{X^*} = \mu_X$, $\Sigma_{XX^*} = \Sigma_{XX}$, and $\Sigma_{ZX^*} = \Sigma_{ZX}$. The closed form estimator for $\widehat{X}$ is then

$$\widehat{X} = \mu_X + \begin{bmatrix} \Sigma_{XX} & \Sigma_{XZ} \end{bmatrix} \begin{bmatrix} \Sigma_{X^*X^*} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{bmatrix}^{-1} \begin{bmatrix} X^* - \mu_X \\ Z - \mu_Z \end{bmatrix}. \tag{2.3.2}$$

This expression is referred to as the BLUP.

All required parameters in Equation (2.3.2) can be consistently estimated in a replicate sample. It is also worth noting that if X and U are both normally distributed, then $\widehat{X}$ is exactly $E[X|X^*, Z]$. Otherwise, if the measurement error variance is sufficiently small, this will provide a reasonable approximation to the mean of X given Z and X^*[10].[2] We can assess the accuracy of this model using standard regression diagnostics. Consider two replicates available, denoted X_1^* and X_2^*. Assuming that $E[X|X_j^*] = \mu + \beta X_j^*$ gives $E[X_1^*|X_2^*] = E[X + U_1|X_2^*] = \mu + \beta X_2^*$. A symmetric argument applies to $E[X_2^*|X_1^*]$. A linear conditional mean, $E[X|X_j^*]$, induces a linear mean in $E[X_j^*|X_l^*]$.

If the mean is sufficiently linear so that $\widehat{X}$ provides a reasonable estimate for the conditional mean of X given $\{X^*, Z\}$, then regression calibration proceeds by conducting

[2] In Chapter 3, we quantify how reasonable this approximation is when normality is not assumed.

the desired analysis, replacing X with $\widehat{X}$. Thus, if we are interested in some parameter Θ, which indexes the distribution of Y given $\{X, Z\}$, expressible as a function $\widehat{\Theta} = \Theta(Y, X, Z)$, then the regression calibration estimator is computed as $\widehat{\Theta}_{\text{RC}} = \Theta(Y, \widehat{X}, Z)$.

Depending on the precise form of $\widehat{\Theta}$, and on the quality of the approximation of $\widehat{X}$ for $E[X|X^*, Z]$, there are some consistency guarantees that can be made. If we assume that the BLUP consistently estimates the conditional mean, and that $\widehat{\Theta}$ is a consistent estimator, then, if $\widehat{\Theta}$ is linear in X, $\widehat{\Theta}_{\text{RC}}$ is unbiased and consistent for Θ [101, 7]. In a log-linear GLM, under certain independence assumptions, the slope parameters will all be consistently estimated but the intercept will not be [7]. In many logistic regression models, the bias of the estimator will be substantially reduced. More importantly, if a logistic regression model is fit with the intention of interpreting the probabilities, rather than the regression coefficients themselves, then the regression calibration correction provides a reasonable approximation. Defining $\text{expit}(x) = (1 + \exp(-x))^{-1}$, then

$$P(Y = 1|X^*, Z) \approx \text{expit}\left\{ \frac{\beta_0 + \beta_X' E[X|X^*, Z] + \beta_Z' Z}{\left[1 + \frac{1}{1.7^2}\beta_X' \text{ var}\left(X|\, X^*, Z\right)\beta_X\right]^{1/2}} \right\},$$

assuming that $P(Y = 1|X, Z) = \text{expit}\left(\beta_0 + \beta_X' X + \beta_Z' Z\right)$ [7].

If the effect size of X or the covariate variance is sufficiently small, then

$$\frac{1}{1.7^2}\beta_X' \text{ var}\left(X|\, X^*, Z\right)\beta_X,$$

is near zero, and the denominator in the approximation tends to 1. If $\widehat{X} = E[X|X^*, Z]$ then the estimated probability will be approximately correct. We refer to regression calibration as *approximately consistent*, with these consistency claims in mind.

Assume that we have n individuals, each with κ_i replicate measurements denoted $X_{i1}^*, \ldots, X_{i\kappa_i}^*$ for $i = 1, \ldots, n$, where each of these measurements are assumed to be independent and identically distributed (iid), coming from a classical additive error model. We also assume that $\{Y_i, Z_i\}$ are measured for all individuals, and are error free. Then, we wish to define $\widehat{X}_i$ as the best linear approximation to $E[X_i|\overline{X}_i^*, Z_i]$, where $\overline{X}_i^* = \kappa_i^{-1}\sum_{j=1}^{\kappa_i} X_{ij}^*$ is the i-th individual sample mean. The following quantities are used as plug-in estimators for Equation (2.3.2). Using these estimators, and Equation (2.3.2), the regression calibration technique computes these quantities, computes $\widehat{X}_i$ for each $i = 1, \ldots, n$, and then uses

$\widehat{\Theta}_{\mathrm{RC}} = \widehat{\Theta}(Y, \widehat{X}, Z)$:

$$\begin{aligned}
\widehat{\Sigma}_{UU} &= \left[\sum_{i=1}^{n}(\kappa_i - 1)\right]^{-1}\left[\sum_{i=1}^{n}\sum_{j=1}^{\kappa_i}(X_{ij}^* - \overline{X}_i^*)(X_{ij}^* - \overline{X}_i^*)'\right]; \\
\widehat{\mu}_X &= \left[\sum_{i=1}^{n}\kappa_i\right]^{-1}\sum_{i=1}^{n}\kappa_i\overline{X}_i^*; \\
\widehat{\mu}_Z &= n^{-1}\sum_{i=1}^{n}Z_i; \\
\nu &= \sum_{i=1}^{n}\kappa_i - \left[\sum_{i=1}^{n}\kappa_i\right]^{-1}\sum_{i=1}^{n}\kappa_i^2; \\
\widehat{\Sigma}_{ZZ} &= (n-1)^{-1}\sum_{i=1}^{n}(Z_i - \overline{Z})(Z_i - \overline{Z})'; \\
\widehat{\Sigma}_{XZ} &= \nu^{-1}\sum_{i=1}^{n}\kappa_i(\overline{X}_i^* - \widehat{\mu}_X)(Z_i - \overline{Z})'; \\
\widehat{\Sigma}_{XX} &= \nu^{-1}\left[\left\{\sum_{i=1}^{n}\kappa_i(\overline{X}_i^* - \widehat{\mu}_X)(\overline{X}_i^* - \widehat{\mu}_X)'\right\} - (n-1)\widehat{\Sigma}_{UU}\right]; \\
\widehat{\Sigma}_{X^*X^*} &= \widehat{\Sigma}_{XX} + \frac{\widehat{\Sigma}_{UU}}{\kappa_i}.
\end{aligned} \tag{2.3.3}$$

2.3.2 Simulation Extrapolation

Simulation extrapolation (SIMEX) is a method that operates by simulating additional measurement error, to determine the impact that this additional error has on the estimators, and then extrapolating this to the case where there is no error at all. Assume that the measurement error variance is known and constant. Our interest is in estimating Θ, through a consistent estimator $\widehat{\Theta}(Y, X, Z)$. To begin, we generate a sequence of random variables $X_{bi}^*(\lambda) = X_i^* + \sqrt{\lambda}\sigma_U\nu_{bi}$, where $\nu_{bi} \overset{iid}{\sim} N(0,1)$ independent of all other quantities, and λ is a given positive constant.

We compute $\widehat{\Theta}_b(\lambda) = \widehat{\Theta}(Y, X_b^*(\lambda), Z)$, and repeat this process across $b = 1, \ldots, B$, for some sufficiently large B. We then define $\widehat{\Theta}(\lambda) = B^{-1}\sum_{b=1}^{B}\widehat{\Theta}(Y, X_b^*(\lambda), Z)$. This serves as a consistent estimator to $E[\widehat{\Theta}_b(\lambda)]$.

This process is repeated over a set of $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_M\}$, generating pairs of observations $\{(\lambda_1, \widehat{\Theta}(\lambda_1)), \ldots, (\lambda_M, \widehat{\Theta}(\lambda_M))\}$. We then posit a functional model for $\widehat{\Theta}(\lambda) = \mathcal{G}(\lambda)$, and call $\mathcal{G}$ the *extrapolant*. Taking the set of estimates, we fit $\mathcal{G}$, giving $\widehat{\mathcal{G}}(\lambda)$. Finally, we compute $\widehat{\Theta}_{\text{SIMEX}} = \widehat{\mathcal{G}}(-1)$. This is sensible since $\text{var}(X^*_{bi}(\lambda)|X) = (1+\lambda)\sigma^2_U$, which takes a value of 0 when $\lambda = -1$. This procedure is known as the *parametric SIMEX* [16, 21].

In the event that σ^2_U is unknown there are two common extensions. The first is to use an estimate of σ^2_U, either computed through the use of replicate measurements (as was the case for regression calibration), or from an external source. Alternatively, if there are iid replicates available, there is a related but modified procedure, called the *empirical SIMEX* [21]. The empirical SIMEX also has the advantage of accommodating heteroscedastic errors across individual observations, a trait which is not shared by the parametric SIMEX.

The empirical SIMEX functions by taking random contrasts of the replicate measures in such a way that these contrasts exhibit the same distributional properties as $X^*_b(\lambda)$. Consider a set of $\mathbf{c}_{b,i} = (c_{b,1}, \ldots, c_{b,\kappa_i})$, such that $\sum_{j=1}^{\kappa_i} c_{b,i,j} = 0$ and $\sum_{j=1}^{\kappa_i} c^2_{b,i,j} = 1$. It can be shown that

$$X^*_{b,i}(\lambda) = \overline{X}^*_i + \sqrt{\frac{\lambda}{\kappa_i}} \sum_{j=1}^{\kappa_i} c_{b,i,j} X^*_{i,j},$$

will exhibit the same moment properties as the previous definition of $X^*_{bi}(\lambda)$. As a result, instead of generating random errors to add to X^*, we could sample random values for $\mathbf{c}$. To do so, one can generate $\nu_{b,i,1}, \ldots, \nu_{b,i,\kappa_i} \overset{iid}{\sim} N(0,1)$, and computing

$$c_{b,i,j} = \frac{\nu_{b,i,j} - \overline{\nu}_{b,i}}{\sqrt{\sum_{j=1}^{\kappa_i} (\nu_{b,i,j} - \overline{\nu}_{b,i})^2}},$$

fulfills the necessary properties [21]. This method can be used in place of the previously discussed method for generating $X^*_{bi}(\lambda)$, and otherwise SIMEX proceeds as described.

Generally, the reliability of either of the SIMEX procedures is determined by the quality of the extrapolant used. One of three forms are typically applied for the extrapolant. All three of these forms will be exact for certain models, assuming the underlying errors are normal [16].

1. The **linear extrapolant**, taking $\mathcal{G}(\lambda) = a + b\lambda$;
2. The **quadratic extrapolant**, taking $\mathcal{G}(\lambda) = a + b\lambda + c\lambda^2$;

3. The **nonlinear extrapolant**[3], taking $\mathcal{G}(\lambda) = a + \frac{b}{c+\lambda}$;

where a, b, and c are the parameters. In many settings the exact form the extrapolant is unknown. In these cases, for $\lambda \geq 0$, the analyst can generate arbitrarily many data points to test the fit of the curve.[4] The nonlinear extrapolant is approximately correct for a large class of models, often working fairly well.

Like regression calibration, SIMEX is an "approximately consistent" correction technique. The SIMEX estimator, under certain assumptions, is consistent for a quantity which will be approximately equal to the true estimator. If we assume that X and U are jointly normal and independent, and that the estimator $\widehat{\Theta}$ is sufficiently smooth [87], then the SIMEX estimator is a consistent estimator whenever $\mathcal{G}$ is correctly specified. If we regard the estimator $\widehat{\Theta}$ as a functional over distributions, denoted $\mathbf{T}$, then we have that $\Theta = \mathbf{T}\left(F_{Y,X,Z}\right)$, where $F_{Y,X,Z}$ specifies the joint distribution of (Y, X, Z).

Taking $*$ to denote the convolution operator, we get that

$$F_{Y,X^*(\lambda),Z} = F_{Y,X,Z} * \Phi_{0,(1+\lambda)\sigma_U^2,0},$$

where $\Phi_{a,b,c}$ denotes the distribution of a normal random variable with mean 0, and covariance matrix with (a, b, c) on the diagonal and zeros elsewhere. Assuming continuity at -1 of $\mathcal{G}(\lambda)$, we get that $\widehat{\Theta}_{\text{SIMEX}} = \widehat{\mathcal{G}}(-1) = \lim_{\lambda \to -1} \widehat{\mathcal{G}}(\lambda)$. If the extrapolant is correctly specified then $\mathcal{G}(\lambda) = \Theta(\lambda) = \mathbf{T}(F_{Y,X,Z} * \Phi_{0,(1+\lambda)\sigma_U^2,0})$, and we get

$$\begin{aligned}
\widehat{\Theta}_{\text{SIMEX}} &\xrightarrow{p} \lim_{\lambda \to -1} \mathcal{G}(\lambda) && \text{(by consistency of the extrapolant estimation)} \\
&= \lim_{\lambda \to -1} \mathbf{T}(F_{Y,X,Z} * \Phi_{0,(1+\lambda)\sigma_U^2,0}) && \text{(by the correctness of the extrapolant)} \\
&= \mathbf{T}\left(\lim_{\lambda \to -1} F_{Y,X,Z} * \Phi_{0,(1+\lambda)\sigma_U^2,0}\right) && \text{(by smoothness or continuity of } \mathbf{T}) \\
&= \mathbf{T}\left(F_{Y,X,Z}\right) && \text{(by normality)} \\
&= \Theta && \text{(by definition of } \mathbf{T}).
\end{aligned}$$

These assumptions are fairly strict in the sense that:

[3]Note the phrasing "nonlinear extrapolant" is taken from Cook and Stefanski [16] and Carroll, Küchenhoff, Lombard, and Stefanski [8]. For clarity, we use the same phrasing. The quadratic extrapolant is of course nonlinear in λ, though the descriptor can instead be taken to referring to the parameters (as-in, this extrapolant cannot be fit with ordinary least squares).

[4]This is limited in practice to the computational feasibility, but, it is an affordance not typically available when fitting models to "real" data.

1. normality is required for the distribution of $X_b^*(\lambda)$ to approach that of X as $\lambda \to -1$;

2. the estimators need to be sufficiently smooth so as to be able to interchange limits; and

3. the extrapolant needs to be correctly specified.

The requirement of normality has been previously discussed; even with an otherwise correct extrapolant, deviations from normality can render the SIMEX estimator unreliable [102, 50]. We will explore the performance of SIMEX more thoroughly, and propose an extension to it which relaxes the need for normally distributed errors, in Chapter 4.

2.3.3 Estimating Equation Approaches

Both regression calibration and simulation extrapolation are attractive owing to their broad utility. These techniques are designed to be applicable to most analyses, and the effects of measurement error are addressed separate from estimation itself.[5] This is an attractive property, particularly in applied settings where regression calibration has seen the largest uptake of any measurement error correction technique [80]. However, as was previously indicated, these methods are generally only approximately consistent. Being designed to be broadly useful has the drawback of these correction strategies not making complete use of all of the information in every problem domain.

Instead of focusing on techniques that work well for a wide variety of estimators, much of the measurement error literature focuses on correcting for the effects of errors in specific settings. These corrections typically require specific mathematical derivations, custom software to fit, and are applicable only in a narrow set of models. This specificity, however, is often associated with improved theoretical properties.

One common approach to developing such methods is through the use of *unbiased estimating equations* (see Yi [101], Carroll, Ruppert, Stefanski, and Crainiceanu [7], and the references therein for a very thorough discussion). Outside of measurement error, unbiased estimating equations (or M-estimators) provide a framework for generalizing likelihood

[5] There is a parallel between these measurement error correction techniques, and imputation for missing data. Imputation corrects for the effects of missing data separate from the estimation procedure, and while it requires certain assumptions about the missingness (and may not always be appropriate), it is widely applicable and correspondingly sees much application.

based estimators[6] [32, 23, 85, 3]. An M-estimator for a parameter, Θ, is the solution, $\widehat{\Theta}$, to the empirical equation

$$n^{-1}\sum_{i=1}^{n}\Psi(\Theta_0; Y_i, X_i, Z_i) = 0,$$

where $E[\Psi(\Theta_0; Y, X, Z)] = 0$ when $\Theta_0 = \Theta$ (the true value). Under regularity conditions, such an estimator $\widehat{\Theta}$ is consistent and asymptotically normal (CAN) for Θ. A more thorough account of unbiased estimating equations is available in Appendix A, where we emphasize the results which our work leverages.

This brief introduction to classical estimation procedures serves to motivate a large class of corrections in the measurement error literature. Unless otherwise specified, we will assume that interest lies in characterizing the conditional distribution of Y given $\{X, Z\}$, parameterized by Θ. We express this conditional density as $f(y|x, z; \Theta)$. We assume that there exists a *conditionally unbiased* estimating function, $\Psi(\cdot)$, for Θ, in the sense that $E[\Psi(\Theta; Y, X, Z)|X, Z] = 0$, when this expectation is taken with respect to the true conditional model. The idea of using estimating equation techniques for measurement error correction is to modify this $\Psi(\cdot)$ in such a way so that it is:

1. computable given the observed data; and
2. unbiased with respect to the (conditional) density that we are working with.

That is, we wish to construct a function Ψ^*, which takes as input $\{Y, X^*, Z\}$, and has the property that, at the true value Θ, $E[\Psi^*(\Theta; Y, X^*, Z)] = 0$. Then, if we solve the empirical estimating equation given by

$$\frac{1}{n}\sum_{i=1}^{n}\Psi^*(\widehat{\Theta}; Y_i, X_i^*, Z_i) = 0,$$

for $\widehat{\Theta}$, under the same standard regularity conditions, $\widehat{\Theta}$ will be a CAN estimator. While this general strategy can be used to derive estimators that are resilient to the effects of measurement error, the details need to be worked through in any specific setting.

In Section 2.5 of Yi [101], several approaches to accomplishing this goal are described in detail. Each of the different techniques is applicable in some settings, depending on the available data, and the form of the estimating equation. While there are various trade-offs

[6] In addition to, for instance, least squares estimators; for our context, thinking of M-estimation as a generalization of likelihood suffices.

between these techniques, for this thesis it suffices to know that they all operate under the same guiding principle. Ultimately, we wish to construct a new estimating function which is computable based on the observed data, and which will be unbiased.

2.3.4 Moment Reconstruction

While the core focus of this thesis will make use of regression calibration, simulation extrapolation, and estimating equation techniques, we also briefly consider *moment reconstruction*. Moment reconstruction is a plug-in based technique, like regression calibration, which provides consistency in a wider class of models [29]. This consistency, however, is achieved through computation that is case dependent, and so it cannot be generally implemented in standard software. Moment reconstruction is similar in spirit to regression calibration, where analysis is conducted using $\widehat{X}_{\mathrm{MR}}$ substituted for X. Here, $\widehat{X}_{\mathrm{MR}}$ is an estimated version of X_{MR} selected such that the joint distribution $(Y, X_{\mathrm{MR}}) \sim (Y, X)$. In general it will be the case that finding X_{MR} such that these distributions match exactly is a challenging problem. However, if instead of exactly matching the distribution we seek to only approximately match the distribution, say, by having the first two joint moments identical, then the problem is made tractable.

In their paper, Freedman et al. [29] prove that

$$X_{\mathrm{MR}}(X^*, Y) = E[X^*|Y](I - G) + X^*G,$$

will match the joint distribution up to the second moment. In this expression,

$$G = \{\mathrm{cov}(X^*|Y)\}^{-1/2} \mathrm{cov}(X|Y)^{1/2},$$

where $A^{1/2}$ denotes the Cholesky decomposition of A. Then, $\widehat{G}(Y)$ is computed based on estimates of these quantities, which will generally depend on the assumed error model, in addition to an estimate for $E[X^*|Y]$. When these quantities are consistently estimated then, asymptotically, $(\widehat{X}_{\mathrm{MR}}, Y)$ will match the first two moments of (X, Y), and intuitively any procedure which relies on only the first two moments of a distribution will perform well with this correction. In the event that the variables are distributed according to a multivariate normal then the joint distribution is entirely specified by the first two moments, and the resultant correction will be consistent.

2.4 Misclassification Models

The discussion throughout this chapter has assumed that the error-prone variable of interest, X, is continuous. If we are instead concerned with a discrete or categorical random quantity, then we will typically refer to this setting as *misclassification*. When a variable of interest is misclassified, this simply means that the observed version, X^*, does not equal the underlying truth X. The same types of categorization relating to measurement error (classical versus Berkson, and differential versus non-differential) can be made in the case of misclassification.

Instead of the structure of error being defined in terms of the distribution of a noise term, U, we instead focus directly on the probability that X is misclassified. Depending on the specific scenario, there are two common ways of modelling the misclassification. We can specify the *misclassification probabilities*, which are defined as

$$P(X^* = x^*|X = x). \tag{2.4.1}$$

Alternatively, a model for the *reclassification probabilities*, given by

$$P(X = x|X^* = x^*), \tag{2.4.2}$$

can be used [101]. The use of these models is analogous to the use of the classical error assumption and the Berkson error assumption, respectively.

In order to decide which framing is more appropriate, it can be helpful to consider the data generation mechanism. Suppose that the true value of X is generated first (say, through the actions of a participant). Then, if X^* is reported afterwards, informed by the truth, using the misclassification probabilities, Equation (2.4.1), is natural since the underlying model generates $X^*|X$. If, on the other hand, the observed value X^* is an antecedent of X, then it is likely more fruitful to consider the reclassification probabilities, Equation (2.4.2), instead. An example of the former scenario would be self-reported smoking status. Here, the true value (X) is based on the patient's underlying behaviour, while the observed value (X^*) is reported after the fact; it is more natural to think of how someone's smoking behaviour would impact their reported smoking behaviour than the reverse. An example of the latter scenario would be related to medication adherence. If a patient is prescribed a particular treatment, this prescription may be recorded as the misclassified response (X^*). Then, the true observation (X) would depend on whether the individual fills their prescription, whether they follow the instructions associated with taking it, and so on. In this case, it is more natural to think about how the prescription (X^*) impacts the adherence behaviour (X).

While some problems lend themselves to one framing or the other, if it is possible to access the marginal models then both framings can be used equivalently. That is because

$$P(X^* = x^*|X = x) = P(X = x|X^* = x^*)\frac{P(X^* = x^*)}{P(X = x)},$$

so long as $P(X = x) \neq 0$. Whether the misclassification is thought of through misclassification or reclassification probabilities, we can also consider the relationship between the outcome and the misclassified variable. We still refer to misclassification where $Y \perp X^*|X$ as *non-differential*, and otherwise we call the misclassification *differential*.

A particularly important case for consideration is when X and X^* are binary. In this case we can specify the misclassification model completely using two values, the *positive predictive value* (PPV) and the *negative predictive value* (NPV).[7] The PPV and NPV are respectively given by $P(X^* = 1|X = 1)$ and $P(X^* = 0|X = 0)$. Here it is equivalent to consider $E[X|X^*] = P(X = 1|X^*)$, as was common in the measurement error setting.

2.5 Correcting for the Effects of Misclassification

While several different correction techniques are relevant for this thesis with respect to measurement error, for misclassification our focus is on estimating equation approaches. The underlying idea for addressing misclassification via estimating equations is equivalent to that with measurement error: an altered estimating equation Ψ^* is found such that $E[\Psi^*(Y, X^*, Z)] = 0$. Just as with measurement error, this is going to typically rely on the specific problem and the underlying assumptions made with regards to the misclassification mechanism.

While our focus will be on unbiased estimating equations, there exist several more general techniques for adjusting for misclassification in a discrete variate.[8] Küchenhoff, Mwalili, and Lesaffre [51] present generalization for the SIMEX technique to account for misclassified variables through a similar procedure: first additional misclassification is simulated, and then the relationship is extrapolated back to the case of no misclassification. Carroll, Ruppert, Stefanski, and Crainiceanu [7] discuss the use of likelihood and quasi-likelihood techniques to account for misclassification in predictors. They also discuss the

[7] We can also use the commonly reported sensitivity and specificity to characterize the model.

[8] Küchenhoff, Mwalili, and Lesaffre [51] observe that "While measurement error models have received much attention in the literature there are only a few papers on misclassification in the context of regression models. This is partly due to the fact that modelling misclassification is in general easier because it is completely characterized by the misclassification matrix [...]"

misclassification of responses, a topic which we do not address in this thesis. Yi [101] demonstrates several examples across different domains (survival analysis, case-control studies, multi-state models, and the misclassification of responses), with a wide variety of techniques to address the issues that arise. The matrix [2] and inverse matrix [53] methods are commonly used and are explored and expanded on by Morrissey and Spiegelman [57].

2.6 Measurement Error in the Example Datasets

While measurement error is a common issue across data from many domains, in this thesis we begin by focusing on the analysis of the FHS and the KLoSA. While both of these studies have previously been analyzed taking into account the error that is present, there are features of the data that are overlooked by common techniques which our proposed methods resolve.

In the FHS, as previously discussed, the primary interest is in the impact of SBP and serum cholesterol on CHD. It has been common to assume that the SBP measurements in the literature are replicate values, subject to a classical additive model. The assumption of replication is violated, in practice, where Carroll, Ruppert, Stefanski, and Crainiceanu [7] note that "... the large-sample test of equality of means has p-value < 0.0001. Thus in fact, the measurement at Exam #2 is not *exactly* a replicate...". Given that our interest is in the long-term average SBP, it seems reasonable to suggest that each clinical measurement taken is a surrogate measurement of the truth. We would like to allow for these measurements to differ in terms of underlying distribution to accommodate this empirical observation. Relaxing this assumption is the focus of Chapter 3.

In the KLoSA, self-reported body weight is used to determine the impact of BMI on an individual's propensity towards being hypertensive. For a small validation sample, in addition to the self-reported body weight measurements, a true clinical measurement was taken. Looking at the distribution of errors it is clear that the errors in the distribution are highly non-normal. Throughout Chapters 3 and 4 we discuss the ways in which common correction techniques rely on the normality assumptions, and in Chapter 4, we propose an adaptation to the SIMEX methodology that relaxes this assumption.

These two datasets serve as illustrative examples for the measurement error problems that we will address. The remainder of the work that follows in Part I presents methods for reducing the impacts of measurement error, in a wide variety of scenarios, while accounting for the types of data that are often observed empirically.

Chapter 3

Generalizations to Measurement Error Models

3.1 Motivation for Error Model Generalizations

A recent survey of applied literature reviewed how researchers tend to address the concerns of measurement error in their analyses [80]. The results of the survey suggest that corrections for the effects of error were rarely applied, and when they were regression calibration was the most widely used correction technique. The study concludes by discussing the need for researchers to be more deliberate in their use of techniques to address the impacts of measurement error, to better discuss the shortcomings of analyses which ignore the impacts of errors, and to leverage the rich literature of methodological advances in measurement error. While these are important considerations from the vantage point of applied researchers, it is also important to consider what is needed from a methodological perspective to assist in the application of techniques which address issues associated with measurement error.

The techniques which are available to correct for the impacts of measurement error are driven largely by what additional data are available. Correction techniques are typically defined with respect to a certain form of auxiliary data (validation samples, replicate measurements, or instrumental data), and often make strong assumptions regarding the exact error mechanism. When the assumptions regarding the auxiliary data or error processes are violated, the application of these correction techniques can result in excess bias compared to a naive analysis [7]. That is to say, it is not the case that "doing anything" is always preferable to "doing nothing". As a result, developing and presenting methods in

a setting which is sufficiently general so as to encapsulate the data which are observed by applied researchers is an important task for those developing methodologies.

It is instructive to note that regression calibration is the most widely used measurement error correction technique observed in the literature. While regression calibration refers to a broad class of techniques, there is a particular focus on the application of regression calibration when replicate measurements are available. In this setting, the standard presentation uses the replicate measurements to estimate the mean and covariance structure associated with the variables of interest, as described in Chapter 2. These estimates are then used to construct the BLUP of the true underlying value. Central to the presentation of the BLUP technique is the assumption that the measurement error terms are independent and identically distributed. Oftentimes the available data would suggest that this is not the case, where instead of the different measurements being identically distributed, each measurement is an error-prone version of the truth.

By way of motivating example we consider the FHS [44]. As noted, the FHS has been analyzed as though measurements of systolic blood pressure and serum cholesterol levels from years prior serve as replicate measurements from today. However, empirically we know that this is not the case. While these differences are small, they suggest that even when the same instrument is being used to measure a quantity of interest, we may be concerned with whether or not the errors are truly identically distributed. This problem is amplified when the different measurements come from entirely different sources.

In nutritional epidemiology, for instance, a researcher may have access to multiple 24 hour recalls (24HRs), which may be analogous to the FHS blood pressure example. It is also possible that, in addition to a 24HR the researchers have access to a food frequency questionnaire (FFQ). In this case while both the 24HR and the FFQ may be seen as error-prone measurements of the truth, it is unlikely that the data from these instruments can be viewed as identically distributed replicate measurements. While regression calibration has been presented using instrumental variables in place of replicate measurements [7], the presentation of these methods is substantially more involved. The integration of both FFQ and multiple 24HRs using regression calibration techniques has also been previously studied [47]. Despite this, the uptake of these correction techniques has been minimal.

Instead of framing correction techniques around replicate measurements, or around instrumental variables generally, we propose a middle ground that is potentially more applicable than replicate measurements and more accessible than instrumental variables. In particular, we propose formulating the error models through the use of *repeated measurements*. Repeated measurements capture any scenario where multiple measurements of the truth are available which are subject to any error processes. This encapsulates the com-

mon setting of independent and identically distributed replicate measurements, but it also captures the setting in the FHS where the error variances shift overtime, or the joint use of FFQs and 24HRs in nutritional epidemiology. The benefit to framing auxiliary data in this way is that it becomes fairly easy for practitioners to assess whether available data are repeated measurements, where it can be challenging to know whether observations are truly replicates, and the conditions for general instruments are fairly opaque.

In addition to assumptions regarding the availability of data, there are typically strong assumptions made regarding the structure of errors. It is common to assume that error-prone measurements of the truth are unbiased, with errors independent of all other terms. While these assumptions are often convenient mathematically, and can be good approximations of the truth, it is important to understand and clearly communicate how violations of these assumptions impact the performance of the corrections.

In nutritional epidemiology, for instance, systematic biases in the errors are common and need to be addressed for valid inference [47]. It will also commonly be the case, for instance with self-reported body weight, that the errors are dependent on the true underlying measurements [100]. These considerations necessarily complicate the applicable methods. Still, the prevalence with which they arise necessitates the development of accessible methods which more correctly account for this complexity, particularly if we wish to see a strong uptake of measurement error corrections in applied literature.

With these considerations in mind, we present a generalized error model. This model is framed around the use of repeated measurements, which may be systematically biased, or exhibit dependency between the errors and underlying truth. Using this framing we demonstrate that, under certain assumptions, common correction techniques (regression calibration through the BLUP and simulation extrapolation) can accommodate a broader range of available data than is typically illustrated in their presentation. Moreover, we show that through simple extensions to these techniques, they can be applied to reduce (or eliminate) bias associated with measurement error in a wide range of settings. Our extensions of these techniques are designed to be accessible, particularly to those familiar with regression calibration or simulation extrapolation. We demonstrate the theoretical validity of the techniques, and illustrate how these ideas can be used for more complex corrections as a means of making theoretically rigorous corrections more broadly applicable.

3.2 Summary of the Proposed Methods

In this chapter we propose a generalized measurement error model centred around the idea of repeated measurements. Suppose that we wish to estimate a parameter θ, which

parameterizes the distribution of $\{Y_i, X_i, Z_i\}$. In place of observing X_i, we observe X^*_{ij}, for $j = 1, \ldots, \kappa_i$ for each individual. Here, X^*_{ij} correspond to a surrogate measurement of the true X_i, perturbed by error. To estimate θ using the observed data, regression calibration and simulation extrapolation are both appealing. Both of these methods assume that X^*_{ij} are identically distributed (meaning specifically that the errors for each surrogate measurement follow the same distribution). If this is not the case then, supposing that κ_i is not the same for all individuals, the standard implementation of these correction techniques no longer consistently corrects the estimators, even when the other assumptions are valid.

Instead, we demonstrate that by allowing each error to be subject to its own distribution, the necessary parameters to perform regression calibration or simulation extrapolation can still be computed. We take the standard estimators for the means, variances, and covariances between the observable quantities (X^*_{ij} and Z_i). Then we use Equation (3.5.2), and Equation (3.5.3) if there are no Z terms and Equation (3.5.4) otherwise, to estimate the moment estimators involving X_i. These can then be used in a standard application of either regression calibration of simulation extrapolation. The application of these techniques can apply to any individual proxy, to a weighted combination of the proxies, or to each individual proxy before combining the resulting estimators.

If one is familiar with regression calibration, or simulation extrapolation, the only necessary change to allow for the accommodation of non-replicated surrogate measurements is by changing from the standard moment estimators for surrounding X_i, Equations (2.3.3), to the moment estimators outlined in this chapter. In addition to discussing the need for these estimators, and their theoretical properties, we also show how they can accommodate measurement error models with bias, or with multiplicative noise. We further demonstrate how these same techniques are applicable outside of regression calibration and simulation extrapolation.

3.3 Generalized Error Structure

The proposed *generalized error structure* assumes that, for each individual $i = 1, \ldots, n$, we observe X^*_{ij} for $j = 1, \ldots, \kappa_i$. Each observation is taken to be an error-prone, repeated measurement. We begin by assuming an additive error structure, and show in Section 3.4 how multiplicative errors can be accommodated as well. For each X^*_{ij}, we suppose that

$$X^*_{ij} = \eta_{0j} + \eta_{1j} X_i + U_{ij} = \eta_{0j} + \eta_{1j} X_i + \delta_j \dot{U}_{ij}. \tag{3.3.1}$$

We assume that U_{ij} is mean-zero, without loss of generality,[1] and that is has variance δ_j^2. This renders $\dot{U}_{ij}$ to be a zero-mean, unit variance random quantity. The error processes are assumed to be independent of one another ($U_{ij} \perp U_{ij'}$ for $j \neq j'$), and we assume that $U_{ij} \perp \{Y_i, X_i, Z_i\}$ for all j, where Y_i is the outcome and Z_i are any additional variates measured without error. We also assumes that all quantities across individuals (for $i \neq i'$) are independent of each other.

This proposed error model accommodates systematic bias through the η_{0j} terms. Despite the assumed independence between U_{ij} and X_i, the use of η_{1j} allows for linear dependence between the error and the truth. To see this note that Equation (3.3.1) can be re-written as $\eta_{0j} + X_i + [(\eta_{1j} - 1)X_i + U_{ij}]$. If we consider $(\eta_{1j} - 1)X_i + U_{ij}$ as the error term in an additive error model, then this error clearly has a linear dependence with X_i, providing a slight relaxation to the assumption of error independence whenever $\eta_{1j} \neq 1$. This error model also makes no assumptions regarding the underlying distribution of U_{ij}, and allows for different variances across the repeated measurements.

While the notation used for the error models seems to imply scalar-valued variates, multivariate random variables can be accommodated through the vectorization of each component. The relevant means and covariance terms are denoted

$$\begin{aligned} E\left[X_j^*\right] &= \eta_{0j} + \eta_{1j} \circ E[X]; \\ \operatorname{cov}\left(X_j^*, X_l^*\right) &= \eta_{1j}^{(d)} \Sigma_{XX} \eta_{1l}^{(d)} + I(j = l) M_j; \\ \operatorname{cov}\left(X, X_j^*\right) &= \Sigma_{XX} \eta_{1j}^{(d)}; \\ \operatorname{cov}\left(Z, X_j^*\right) &= \Sigma_{ZX} \eta_{1j}^{(d)}. \end{aligned} \tag{3.3.2}$$

Additionally, we define $E\left[X_j^* \middle| X\right] = \eta_{0j} + \eta_{1j} \circ X$ and $\operatorname{var}\left(X_j^* \middle| X\right) = M_j(X)$. Here, $\circ$ represents the Hadamard (element-wise) product, and $\eta_{1j}^{(d)}$ represents the diagonal matrix with the elements of η_{1j} along its diagonal. M_j and $M_j(X)$ are matrices that will capture the variance of the assumed error model, taking the form of $M_j = M_j(X) = \Sigma_{U_j U_j}$ for the additive structure that we have assumed. If we have a vector $\nu = (\nu_1, \cdots, \nu_p)'$, then we define the inverse vector $\nu^{-1} = (1/\nu_1, \cdots, 1/\nu_p)'$.

[1] If U_{ij} were not mean zero, it could be centred, and its mean could be absorbed into η_{0j}.

3.4 Multiplicative Measurement Error

Much of the existing literature on measurement error corrections assumes an additive structure for the measurement error model. Eckert, Carroll, and Wang [24] convincingly argue that, owing in part to the large existing literature, it is advisable to transform measurements to a scale where the errors are additive. They go on to propose methods of finding transformations which allow for the recovery of an additive structure. Their methods allow for the estimation of a function h such that $h(X^*) = h(X) + U$, for a U which is independent of X. Moreover, they derive these transformations without making any distributional assumptions regarding X.

As a general rule, we recommend following the advice of these authors, and searching for transformations to additivity whenever possible. This permits access to the rich literature of measurement error correction techniques. Moreover, their proposed methods can find transformations which render the error distribution known. However, in the context of repeated (rather than replicated) measurements, it is possible that one proxy measurement has errors on an additive scale, and another which has errors that would require a transformation. Moreover, it may not be known a priori whether a particular instrument is likely to be additive or not.

The proposed error models are readily generalized to allow for a multiplicative error structure. This framework accommodates

$$X^*_{ij} = \eta_{0j} + \eta_{1j} X_i V_{ij} = \eta_{0j} + \eta_{1j} X_i (1 + \delta_j \dot{U}_{ij}). \tag{3.4.1}$$

Here, $\eta_{0j}, \eta_{1j}, \delta_j$ and $\dot{U}_{ij}$ are as in Equation (3.3.1). This can be made multivariate using Hadamard products in place of scalar multiplication. The presented mean and covariance structure in Equations (3.3.2) apply in this model, where

$$M_j = \eta_{1j}^{(d)} \left(E[XX'] \circ \Sigma_{V_j V_j} \right) \eta_{1j}^{(d)},$$

and

$$M_j(X) = \eta_{1j}^{(d)} \left(XX' \circ \Sigma_{V_j V_j} \right) \eta_{1j}^{(d)}.$$

If it is suspected that many, or most, of the available repeated measurements follow a non-additive structure, then use of transformations is advised. However, the generality of the proposed measurement error model means that corrections within this general framework can make use of variates that are measured with multiplicative error. Moreover, the analyst need not specify whether a particular proxy is subject to additive or multiplicative error within this framework.

3.5 Parameter Identification

To motivate the necessary parameters in the model, we consider standard regression calibration. Regression calibration through the use of the BLUP proceeds on the basis of an assumed linear model for $E[X|X^*, Z]$, where X^* is a combination of replicate measurements (typically, the sample mean). If we suppose that, for all individuals i, $\kappa_i = k$ and we define $X_i^* = \frac{1}{k}\sum_{j=1}^{k} X_{ij}^*$, then we can consider the relationship between X_i and $\{X_i^*, Z_i\}$. Specifically, we can make the standard BLUP argument, where we consider linear estimators of X, and take $\widehat{X} = \mu + \beta X^* + \gamma Z$, such that $E[(X - \widehat{X})'(X - \widehat{X})]$ is minimized. The closed form estimator for $\widehat{X}$ is given by

$$\widehat{X} = \mu_X + \begin{bmatrix}\Sigma_{XX^*} & \Sigma_{XZ}\end{bmatrix} \begin{bmatrix}\Sigma_{X^*X^*} & \Sigma_{X^*Z} \\ \Sigma_{ZX^*} & \Sigma_{ZZ}\end{bmatrix}^{-1} \begin{bmatrix}X^* - \mu_{X^*} \\ Z - \mu_Z\end{bmatrix}. \tag{3.5.1}$$

Suppose that, in addition to complete replication, all X_{ij}^* are subject to additive noise. Then we can view X_i^* as an error-prone measurement itself, given by

$$X_i^* = \frac{1}{k}\sum_{j=1}^{k} \eta_{0j} + \left(\frac{1}{k}\sum_{j=1}^{k} \eta_{1j}\right) X_i + \frac{1}{k}\sum_{j=1}^{k} U_{ij}.$$

If we further assume that $\eta_{0j} = 0$ for all j and that $\eta_{1j} = 1$ for all j, then this simplifies to

$$X_i^* = X_i + U_i,$$

where $U_i = \frac{1}{k}\sum_{j=1}^{k} U_{ij}$ has mean zero and variance $\frac{1}{k^2}\sum_{j=1}^{k} \delta_j^2$. In this setting it is straightforward to show that the standard estimator, $\widehat{\Sigma}_{UU}$, is consistent for $\frac{1}{k}\sum_{j=1}^{k} \delta_j^2$, and as a result, $\widehat{\Sigma}_{X^*X^*} = \widehat{\Sigma}_{XX} + \widehat{\Sigma}_{UU}/k$ will consistently estimate the variance of X_i^*. Taken together, this renders the standard BLUP-based regression calibration a valid correction technique in the general model, so long as every individual has k unbiased, repeated measurements.

Suppose that instead we observe partial replication, such that $1 \leq \kappa_i \leq k$ is not constant across all individuals. In this case, defining $X_i^* = \frac{1}{\kappa_i}\sum_{j=1}^{\kappa_i} X_{ij}^*$ renders X_i^* to be non-identically distributed across different individuals, even when all measurements remain unbiased proxies. The concern is that

$$\text{var}(X_i^*) = \frac{1}{\kappa_i^2}\sum_{j=1}^{k} r_{ij}\delta_j^2,$$

where $r_{ij} = 1$ if X^*_{ij} is observed, and is zero otherwise. As a result, the set of observed proxies for each individual dictates the variance of X^*_{ij}. The estimator $\widehat{\Sigma}_{UU}$ will remain consistent for $\frac{1}{k}\sum_{j=1}^{k}\delta_j^2$, however, this quantity cannot be directly transformed into $\text{var}(X^*_{ij})$. Thus, under the assumption of non-complete replication, with non-identically distributed repeated measurements, the standard regression calibration procedure fails to produce meaningful estimates of the required correction parameters.

If some of the proxies are biased, either with $\eta_{0j} \neq 0$ or $\eta_{1j} \neq 1$, then regardless of whether there is complete replication or not, the standard regression calibration correction will be invalid, since the simplifying assumption that $\mu_{X^*} = \mu_X$ will not hold. Considering Equation (3.5.1), taking X^* to be a single error-prone proxy, we see that supposing this model is linear, then we require an estimate for the means and covariances of $\{X, X^*_j, Z\}$. This motivates the derivation of specific estimators for each of these moment quantities. Like regression calibration, many existing measurement error correction techniques rely on these types of moment estimators: any such correction is amenable to the proposed generalized measurement error correction technique, through the following parameter estimators.

While standard estimators exist for all of the observable components,

$$\{\mu_{X^*_j}, \Sigma_{X^*_j X^*_j}, \mu_Z, \Sigma_{ZZ}, \Sigma_{X^*_j Z}\},$$

we specifically require the ability to identify μ_X, $\Sigma_{XX^*_j}$, and Σ_{ZX}. In full generality, this assumed model structure will lead to identifiability concerns. We must impose restrictions on some model parameters in order to render the parameters estimable. We will assume that, for some known set of j, (1) $\eta_{0j} = 0$, (2) $\eta_{1j} = 1$, or (3) both $\eta_{0j} = 0$ and $\eta_{1j} = 1$. These assumptions will also capture the case where, for instance, $\eta_{0j} = c$ for any known constant c. If c is non-zero, we can take $X^*_j - c$, leaving us with a measurement satisfying assumption (1). When $\eta_{0j} = 0$ and $\eta_{1j} = 1$ for all j, our model reduces to that of having κ_i unbiased measurements of X, from possibly different distributions.

For notational convenience, we define J_0, J_1, and J_{01} to be the index sets for the proxies corresponding to assumptions (1), (2), and (3) respectively. In this notation, $J_{01} = J_0 \cap J_1$. We will assume that $|J_0| \geq 1$ and $|J_1| > 1$. These assumptions will suffice for the identification of the parameters, but are not strictly necessary. Under these assumptions, we take

$$\widehat{\mu}_X = \frac{1}{|J_0|}\left[\sum_{j\in J_{01}} \widehat{\mu}_{X^*_j} + \sum_{j\in J_0\setminus J_{01}} \widehat{\eta}_{1j}^{(d)^{-1}}\widehat{\mu}_{X^*_j}\right]. \tag{3.5.2}$$

If Z is not observable, then we can take

$$\begin{aligned}\widehat{\Sigma}_{XX_j^*} &= \frac{1}{|J_1 \setminus \{j\}|} \sum_{l \in J_1 \setminus \{j\}} \widehat{\Sigma}_{X_l^* X_j^*}; \\ \widehat{\eta}_{1j}^{(d)} &= \frac{1}{k-1} \sum_{l \neq j; l=1}^{k} \widehat{\Sigma}_{X_j^* X_l^*} \widehat{\Sigma}_{XX_l^*}^{-1}.\end{aligned} \tag{3.5.3}$$

If Z is observable then $|J_1| = 1$ is permissible and we take

$$\begin{aligned}\widehat{\Sigma}_{ZX} &= \frac{1}{|J_1|} \sum_{j \in J_1} \widehat{\Sigma}_{ZX_j^*}; \\ \widehat{\eta}_{1j}^{(d)} &= \left\{ \left(\widehat{\Sigma}_{XZ} \widehat{\Sigma}_{ZX} \right)^{-1} \widehat{\Sigma}_{X_j^* Z} \widehat{\Sigma}_{ZX_j^*} \right\}^{1/2}; \\ \widehat{\Sigma}_{XX_j^*} &= \frac{1}{k-1} \sum_{l \neq j; l=1}^{K} \widehat{\eta}_{1l}^{(d)^{-1}} \widehat{\Sigma}_{X_l^* X_j^*}.\end{aligned} \tag{3.5.4}$$

Lemma 3.5.1 (Estimating Equations for Parameters)**.** *Take ξ to be the vector of moment parameters for $\{X, X_1^*, \ldots, X_k^*, Z\}$. Then, under regularity conditions, for g given as Equation (B.1.1), the estimator $\widehat{\xi}$ that solves $n^{-1} \sum_{i=1}^{n} g(X_i^*, Z_i, \xi) = 0$, is consistent and asymptotically normal for the true ξ. As $n \to \infty$,*

$$\sqrt{n} \left(\widehat{\xi} - \xi \right) \xrightarrow{d} N \left(\mathbf{0}, \mathcal{A}^{-1}(\xi) \, \mathcal{B}(\xi) \, \mathcal{A}^{-1}(\xi)' \right),$$

where $\mathcal{A}(\xi) = E \left[\frac{\partial}{\partial \xi'} g(X^*, Z, \xi) \right]$ and $\mathcal{B}(\xi) = E \left[g(X^*, Z, \xi) g(X^*, Z, \xi)' \right]$.

Note that in order to consistently estimate these parameters when we have incomplete replication, we have to assume that the set of replicates available for each individual are independent of the measured variables. Under this assumption of *ignorable missingness*, the j-th proxy's parameters are computable consistently using only the observations which have the j-th proxy available. The function that the M-estimators are based on, $g(\cdot)$, can be simply modified to include observation indicators of X_{ij}^*.

Lemma 3.5.1 is particularly useful when we consider measurement error correction techniques that rely upon the moment estimators contained in ξ through the use of additional M-estimators. Regression calibration and simulation extrapolation are both such techniques. We can derive the asymptotic distribution for any correction technique which can

be framed as an M-estimator involving the parameters in ξ.

Lemma 3.5.2 (Asymptotic Distribution of $\widehat{\Theta}$)**.** *Assume that $\widehat{\xi}$ solves the empirical estimating equation from Lemma 3.5.1, denoted $g_n(\widehat{\xi}) = 0$, and that $\widehat{\Theta}$ is a solution to the empirical estimating equation $U_n(\Theta, \widehat{\xi}) = 0$, where $\widehat{\xi}$ and $\widehat{\Theta}$ are estimating ξ and Θ, respectively. Then we have that*

$$\sqrt{n}\left(\widehat{\Theta} - \Theta\right) \xrightarrow{d} N\left(\mathbf{0}, \Sigma_{(1)}\right),$$

as $n \to \infty$, where $\Sigma_{(1)} = Q\,\mathcal{A}^{-1}(\Theta, \xi)\,\mathcal{B}(\Theta, \xi)A^{-1}(\Theta, \xi)'Q'$, for $Q = \begin{bmatrix} I_{p\times p} & 0_{p\times q} \end{bmatrix}$, $\mathcal{A}(\Theta, \xi)$ is upper-triangular, and $\mathcal{B}(\Theta, \xi)$ is symmetric. Here p is the dimension of Θ, and q is the dimension of ξ.

3.6 Generalization of Regression Calibration

Using the previous discussion, the BLUP based on each X_j^* can be computed, by taking $X^* = X_j^*$ in Equation (3.5.1), and applying Lemma 3.5.1. Doing this separately across all proxies is not likely to make efficient use of the observed data. Instead, one of two options can be considered. The first is to define X^* to be a weighted combination of X_j^*. That is,

$$X_i^* = \frac{\sum_{j=1}^k r_{ij}\alpha_j X_{ij}^*}{\sum_{j=1}^k r_{ij}\alpha_j},$$

where $\sum_{j=1}^k \alpha_j = 1$. Taking this approach with $\alpha_j = 1/k$ for all $j = 1, \ldots, k$ results in the standard correction based on the mean, as discussed in the previous section. In the case of non-identically distributed measurements, it is unlikely that this will be the most efficient combination. Intuitively, the measures with lower measurement error variance ought to contribute more to our proxy measure than those with higher measurement error variance.

Instead, the parameters α can be added to $\{\mu, \beta, \gamma\}$, and then determined during the minimization of the BLUP. That is, we can find the parameters $\{\mu, \beta, \gamma, \alpha\}$ that minimize $E[(\widehat{X}_i - X_i)'(\widehat{X}_i - X_i)]$. This provides the set of optimal weights in the same sense that the BLUP provides the optimal choice of $\widehat{X}$. The downside to this technique is that it will not be possible, in general, to derive a closed form expression for the set of weights.

Alternatively, we can view the problem of generating estimates of $\widehat{X}_i$ through an ex-

pression given by

$$\mu + \sum_{j=1}^{k} \beta_j X_{ij}^* + \gamma Z_i.$$

In the above methods we take $\beta_\ell = 0$ for all $\ell \neq j$ when we wish to use a single proxy, and take $\beta_j = \alpha_j \beta$ when using a weighted combination. This framing, while perhaps less natural based on the standard regression calibration procedure, provides a more direct method for estimating the BLUP for each individual. Working through the standard BLUP calculations presents an estimator that is given by

$$\widehat{X} = \mu_X + \begin{bmatrix} \Sigma_{XX_1^*} & \cdots & \Sigma_{XX_K^*} & \Sigma_{XZ} \end{bmatrix} \begin{bmatrix} \Sigma_{X_1^*X_1^*} & \cdots & \Sigma_{X_1^*X_K^*} & \Sigma_{X_1^*Z} \\ \vdots & \ddots & \vdots & \vdots \\ \Sigma_{X_K^*X_1^*} & \cdots & \Sigma_{X_K^*X_K^*} & \Sigma_{X_K^*Z} \\ \Sigma_{ZX_1^*} & \cdots & \Sigma_{ZX_K^*Z} & \Sigma_{ZZ} \end{bmatrix}^{-1} \begin{bmatrix} X_1^* - \mu_{X_1^*} \\ \vdots \\ X_K^* - \mu_{X_K^*} \\ Z - \mu_Z \end{bmatrix}.$$

One advantage to this framing is that it immediately becomes clear how to handle incomplete replication. The same argument can be applied to an individual with a subset of the available observations, giving an equivalent form without the unobserved terms. Then, for each individual the BLUP can be estimated as $\widehat{X}_i$, and these imputed values can then be used in place of X_i in the analysis.

We refer to these two strategies as *combining proxies* or *combining estimating equations*, respectively. It is our belief that for practitioners it will be more familiar to combine proxies directly, as it more closely resembles extant methods. However, we note that computing the optimal weights can be numerically unstable depending on the observed data. This technique is implemented through the use of numerical optimization, and in many situations the differences in estimator efficiency will be small. In practice, an analyst wishing to use the proxy combination strategy can likely equally weight proxies for each individual, supposing that they appear roughly in line with one another (as is frequently the case). The combination of estimating equations, while perhaps less familiar, does not exhibit the same concerns as working through the optimal weights (as it is available in a closed form). Correspondingly, this technique, may be easier to implement in practice.

Any of these strategies can be used to compute the parameters necessary to estimate the BLUP for each individual. Using this estimated BLUP, we can then take $\widehat{\Theta}_{\text{RC}}$ to the be solution to $U_n(Y, \widehat{X}, Z; \widehat{\Theta}_{\text{RC}}) = 0$, where Θ is the parameter of interest. Conceptually, this strategy is no different from the standard case. We model $\widehat{X} = E[X|X^*, Z]$ as a linear function, and then use the estimated value in place of the truth. The core distinction

between the standard implementation and ours is a recognition that, when X_j^* are not identically distributed, when they are not universally observed, or when they may exhibit systematic bias, an equal weighting combination ranges from inefficient to inconsistent. By modelling each separately, sharing parameters only when appropriate, the same correction strategy becomes applicable to a far wider range of problems.

When using the BLUP, the conditional expectation is consistently estimated only when it is linear in the conditioning variables. In Appendix B, Lemma B.1.1 is provided as a generalization of Lemma A.1 from Carroll and Stefanski [10]. It uses their notation for matrix derivatives and the trace operator. Using this Lemma, we can characterize the linearity of the BLUP.

Theorem 3.6.1 (General Form of Conditional Means)**.** *Under the generalized error models presented, assuming that* $E[U|X] = 0$*, and denoting* $\mathrm{cov}(U|X = x) = \Omega(x)$ *and the density of* X^* *as* $f_{X^*}(x)$*, we have that*

$$E[X|X^*] = \eta_1^{-1}\left\{X^* - \eta_0 + \delta^2\left[\mathrm{Tr}\left(\frac{\partial}{\partial x}\Omega(x)\right) + \Omega(x)\frac{f'_{X^*}(x)}{f_{X^*}(x)}\right]_{x=X^*}\right\} + O_p(\delta^3),$$

when $X^* = \eta_0 + \eta_1 X + \delta U$ *and*

$$E[X|X^*] = \eta_1^{-1}\left\{1 + \delta^2\left[2\cdot\mathrm{diag}\left(\Omega(x)\right) + x\circ\left(\mathrm{Tr}\left(\frac{\partial\Omega(x)}{\partial x}\right) + \Omega(x)\frac{f'_{X^*}(x+\eta_0)}{f_{X^*}(x+\eta_0)}\right)\right]_{x=X^*-\eta_0}\right\}(X^* - \eta_0) + O_p(\delta^3),$$

when $X^* = \eta_0 + \eta_1 X(\mathbf{1} + \delta U)$.

The term $f'_{X^*}(x)/f_{X^*}(x)$ is linear in x if and only if $X^* \sim N(\mu, \sigma^2)$ [10]. Since we are conditioning on X^*, we can exclude values of this ratio which are unobservable almost surely. As a result, domain indicators can be dropped. Then, for the case of additive errors, our conditional mean will be approximately linear if either $\Omega(X^*)$ is linear and $f'_{X^*}(x)/f_{X^*}(x)$ is constant, or if $\Omega(X^*)$ is constant and $f'_{X^*}(x)/f_{X^*}(x)$ is linear. Linearity in the multiplicative case is more restrictive. Here, due to the additional multiplicative X^* term, we need both $\mathrm{diag}(\Omega(X^* - \eta_0))$ to be constant and

$$\mathrm{Tr}\left(\frac{\partial\Omega(x)}{\partial x}\right) + \Omega(x)\frac{f'_{X^*}(x+\eta_0)}{f_{X^*}(x+\eta_0)} = 0. \tag{3.6.1}$$

If $\Omega(x)$ is constant, then the first term in Equation (3.6.1) will be 0. In order for the second term in this expression, to be 0 we would either require that $\mathrm{cov}(U|X) = 0$ or that $f_{X^*}(x)$

is constant. As a result, it is sufficient to have $X^* \sim \text{Unif}(\cdot)$, and for $\Omega(x)$ to be constant. This illustrates the caveats with applying this method to multiplicative errors. In many situations, the linear approximation for the additive case will be sufficiently good so as to achieve the near consistent results that are often claimed. However, in the multiplicative case, the expectation will be non-linear under most assumed models.

Direct calculations show that, in order for $E[X|X_j^*]$ to be approximately linear, additive models require $E[U_j|X_j^*]$ to be linear, and multiplicative models require $E[(1+\delta_l U_l)^{-1}|X_l^*]$ to be constant. Consider two observations under the assumed models. Denoting an additive surrogate measurement as $X_j^{(A*)}$ and a multiplicative one as $X_l^{(M*)}$, then

$$
\begin{aligned}
E\left[X_1^{(A*)}\middle| X_2^{(A*)}\right] &= \eta_{01} + \eta_{11} \circ \eta_{12}^{-1} \circ \left\{X_2^{(A*)} - \eta_{02} - \delta_2 E\left[U_2 \middle| X_2^{(A*)}\right]\right\}; \\
E\left[X_1^{(A*)}\middle| X_2^{(M*)}\right] &= \eta_{01} + \eta_{11} \circ \eta_{12}^{-1} \circ \left\{X_2^{(M*)} - \eta_{02}\right\} \circ E\left[(1+\delta_2 U_2)^{-1}\middle| X_2^{(M*)}\right]; \\
E\left[X_2^{(M*)}\middle| X_1^{(A*)}\right] &= \eta_{02} + \eta_{12} \circ \eta_{11}^{-1} \circ \left\{X_1^{(A*)} - \eta_{01} - \delta_1 E\left[U_1 \middle| X_1^{(A*)}\right]\right\}; \\
E\left[X_2^{(M*)}\middle| X_1^{(M*)}\right] &= \eta_{02} + \eta_{12} \circ \eta_{11}^{-1} \circ \left\{X_1^{(M*)} - \eta_{01}\right\} \circ E\left[(1+\delta_1 U_1)^{-1}\middle| X_1^{(M*)}\right].
\end{aligned}
$$

When conditioning on $X_j^{(A*)}$, we see that if $E[U_j|X_j^{(A*)}]$ is linear in X_j^* then these conditional expectations simplify to a linear function. Similarly, when conditioning on $X_l^{(M*)}$, we see that if $E[(1+\delta_l U_l)^{-1}|X_l^{(M*)}]$ is constant then these simplify to be linear. Checking the goodness of fit of a linear model between any two proxies in turn checks the ability of $E[X|X_j^*]$ to be approximated by a linear model. This also highlights the relation between our methodology and the standard instrumental variable approaches, which are based on regressing a measurement of the truth on an instrument [7].

These results justify both the theoretical conditions under which a linear model is warranted, and a mechanism for checking whether or not linearity holds approximately. If linearity approximately holds then the modified regression calibration procedure may be warranted, and the resultant estimators will be asymptotically normal.

Theorem 3.6.2 (Asymptotic Normality of Regression Calibration)**.** *Under regularity conditions, the estimator $\widehat{\Theta}_{RC}$ is consistent for Θ_{RC}, and is asymptotically normally distributed, such that as $n \to \infty$,*

$$\sqrt{n}\left(\widehat{\Theta}_{RC} - \Theta_{RC}\right) \xrightarrow{d} N\left(\mathbf{0}, \Sigma_{RC}\right),$$

where $\Sigma_{RC} = Q\,\mathcal{A}_{RC}^{-1}\mathcal{B}_{RC}\,\mathcal{A}_{RC}^{-1'}\,Q'$, for matrices analogous to those in Lemma 3.5.2.

Importantly, this result shows asymptotic normality, not around the true values for

the parameter Θ, but rather around Θ_{RC}, the solution to $U(Y, X_{\text{RC}}, Z, \Theta_{RC}) = 0$. Here X_{RC} is the *true* BLUP, which may differ from the true conditional mean. Under regularity conditions, Θ_{RC} is the probability limit of $\widehat{\Theta}_{\text{RC}}$. The asymptotic performance is determined by the difference between Θ and Θ_{RC}. Consistency is achieved when $\Theta = \Theta_{\text{RC}}$.

In the standard setting, the BLUP based regression calibration estimators are consistent only when both the true BLUP is $E[X|X^*]$, and a linear model for $E[Y|X]$ is valid. Results regarding consistency, and approximate consistency, of regression calibration methods generally will apply to the modified technique, under the caveat that these are derived when $\widehat{X} \xrightarrow{p} E[X|X^*]$, as $n \to \infty$.

3.7 Generalization of SIMEX

We rely on M_j and $M_j(X)$ to motivate the modified versions of SIMEX. The strategy is to match the first two moments of X and $X_b^*(\lambda)$, if $\lambda = -1$. For fixed $\lambda \geq 0$, take

$$X_{b,j}^*(\lambda) = \eta_{1j}^{-1} \circ \left\{ X_j^* - \eta_{0j} + \sqrt{\lambda} M_j^{1/2} \nu_{bj} \right\}, \tag{3.7.1}$$

where ν_{bj} is an appropriately sized standard normal pseudo-random variable, independent of all covariates. Given X, we find that $E[X_b^*(\lambda)|X] = \eta_{1j}^{-1} \circ \{\eta_{0j} + \eta_{1j} \circ X - \eta_{0j}\} = X$. Similarly, $\text{cov}\left(X_b^* | X\right) = (1+\lambda)\eta_{1j}^{(d)-1} M_j(X) \eta_{1j}^{(d)-1}$. As a result, $X_b^*(\lambda)$ agrees with X up to the second moment, as $\lambda \to -1$.

As in the standard SIMEX, we do not typically have M_j or $\eta_{\cdot}$ available, and as a result we will estimate them from the proxy observations. While in Section 3.5 we did not explicitly write down estimators for η_{0j} or M_j, both of these can be obtained as simple transformations for quantities which are estimated in that estimating equation. We can re-write $X_{b,j}^*(\lambda)$ as

$$X_{b,j}^*(\lambda) = \eta_{1j}^{-1} \circ \left\{ X_j^* - (\mu_{X_j^*} - \eta_{1,j} \circ \mu_X) + \sqrt{\lambda} (\Sigma_{X_j^* X_j^*} - \eta_{1j}^{(d)} \Sigma_{XX} \eta_{1j}^{(d)})^{1/2} \nu_{bj} \right\}.$$

As a result, we can still make use of Lemma 3.5.1 in the context of simulation extrapolation.

This raises a question regarding how to best implement the method, taking into account the proxies. In the standard case, if homoscedasticity is assumed, then SIMEX progresses using $\overline{X^*}$ and $\widehat{\Sigma}_{UU}$. As discussed with standard regression calibration, if $\kappa_i = k$, the standard SIMEX applies to non-iid data, with the same caveats: namely, if there is complete replication and all of the proxies are unbiased.

Just as with regression calibration, there are two natural ways of making use of all X_j^* in the correction technique, combining proxies or combining estimating equations. To combine the proxies we use Equation (3.7.1), where in place of X_j^* we take $X^* = \sum_{j=1}^k \alpha_j X_j^*$ for some set of weights α. Unlike in the case of the BLUP where the weights could be selected through the minimization of the MSE, a specific objective function would be required in this setting. One sensible option is to use inverse variance weighting, providing X^* which has the minimal variance among all linear combinations. This is an appealing choice as the weights are available in closed form, and have an analogue to the standard SIMEX.

Combining the estimating equations is less intuitive than in the regression calibration setting. One technique for doing this would be to use each individual proxy to estimate the parameter of interest, which gives us several different estimators for the true parameter, say $\widehat{\Theta}_{\text{SIMEX,j}}$ for each $j = 1, \ldots, k$. Then, we can combine these k estimates directly. That is, taking $\widehat{\Theta}_{\text{SIMEX}} = \sum_{j=1}^k \zeta_j \widehat{\Theta}_{\text{SIMEX,j}}$. If $k = 2$, then the optimal weights would be

$$\zeta_j = \frac{1}{2} + \left\{ 4 \operatorname{cov}\left(\widehat{\Theta}_{\text{SIMEX}}^{(1)}, \widehat{\Theta}_{\text{SIMEX}}^{(2)} \right) \right\}^{-1} \left[\operatorname{var}\left(\widehat{\Theta}_{\text{SIMEX}}^{(j)} \right) - \operatorname{var}\left(\widehat{\Theta}_{\text{SIMEX}}^{(\ell)} \right) \right].$$

These weights require estimates for the variance and covariance of the different estimators which can be quite computationally intensive. In simulations, this strategy of combining distinct estimators itself performed notably worse than the strategy of first combining proxies, and as a result is not generally recommended at this point.[2]

The modified SIMEX correction is approximately consistent in the same way as the standard SIMEX. Viewing the extrapolant as a functional on distributions, $\widehat{\Theta}_{\text{SIMEX}}$ will be consistent for $\lim_{\lambda \to -1} \mathcal{G}(X_b^*(\lambda))$, which we call Θ_{SIMEX}. The SIMEX estimators will generally be asymptotically normal.

Theorem 3.7.1 (Asymptotic Normality of SIMEX)**.** *Under regularity conditions, the estimated parameters using the SIMEX correction, $\widehat{\Theta}_{SIMEX}$ are consistent for the parameters Θ_{SIMEX}, and are asymptotically normally distributed, such that*

$$\sqrt{n}\left(\widehat{\Theta}_{SIMEX} - \Theta_{SIMEX} \right) \xrightarrow{d} N(\mathbf{0}, \Sigma_{SIMEX}),$$

as $n \to \infty$, where Σ_{SIMEX} is estimable through sandwich estimation techniques.

[2]To apply this strategy in simulation, however, is a substantial computational burden; as a result, we may see improved performance had better estimates of the relevant variance terms been used.

3.8 Simulation Studies

To investigate the behaviour of the proposed methods, we consider three simulated scenarios. Our simulations compare our proposed estimators with the standard implementation of these techniques, using the generalized error models, across a variety of settings for which SIMEX and regression calibration are known to be effective.

3.8.1 Linear Regression Models

In the first simulation we consider a linear regression. We take $X = \begin{bmatrix} X_1 & X_2 & X_3 \end{bmatrix}$, with $X_1 \sim N(0,1)$, $X_2 \sim N(3,2)$, and $X_3 \sim N(1,3)$ to be the true covariate vector, where all components are assumed to be independent. The outcome is taken to be $Y = 2 - X_1 + 2X_2 + 0.5X_3 + \epsilon$, where $\epsilon \sim N(0,1)$. We generate three proxies, $X_1^* = X + \begin{bmatrix} U_{11} & U_{12} & U_{13} \end{bmatrix}$, where $U_{1j} \overset{iid}{\sim} N(0,1)$, $X_2^* = X + \begin{bmatrix} U_{21} & U_{22} & U_{23} \end{bmatrix}$ with $U_{21} \sim N(0,1)$, $U_{22} \sim N(0,4)$ and $U_{23} \sim N(0,3)$, and $X_3^* = X + \begin{bmatrix} U_{31} & U_{32} & U_{33} \end{bmatrix}$ where $U_{31} \sim N(0,2)$, $U_{32} \sim N(0,2)$ and $U_{33} \sim N(0,5)$. We select 50% of X_2^* and 20% of X_3^* to be missing.

We estimate model parameters using (1) standard regression calibration, (2) standard SIMEX, (3) empirical SIMEX, (4) generalized regression calibration using fixed weights, (5) generalized regression calibration solving for optimal weights, (6) generalized SIMEX where proxies are combined, and (7) generalized SIMEX where the estimates are combined. These simulations were repeated 1000 times with a sample size of 5000. The results for all scenarios are included in Table 3.1.

From the results in the table we see that the standard regression calibration and standard SIMEX estimators are outperformed (in bias and MSE) by the generalized versions (both weighted and not). In this example, it is worth considering the performance of the empirical SIMEX. This method tended to perform slightly better than the generalized regression calibration throughout. This can be explained through the fact that all errors are normally distributed and, as a result, combinations of the errors also remain normal, which is the critical requirement for the implementation of the empirical SIMEX.

3.8.2 Log-Linear Regression Models

Next, we consider a log-linear model. We generate $Z \sim \text{Binom}(0.3)$ and $X \sim N(0.02Z, 0.5)$. The outcome is a gamma random variable such that $E[Y|X,Z] = \exp(2 - 3Z + 2X)$. We generate three proxies, with $X_1^* = XV_1$ where $V_1 \sim \text{Unif}(0.7, 1.3)$, and $X_2^*, X_3^* \overset{iid}{\sim} X +$

Table 3.1: Results from a simulation of a linear regression model comparing the generalized and standard techniques. The tables show the MSE, mean, and median bias for the intercept and slope parameters, estimated in 1000 ($n = 5000$) replicated simulations.

			Bias	
	Method	MSE	Mean	Median
Intercept	Standard Regression Calibration	0.334	0.562	0.555
	Standard SIMEX	4.608	2.094	2.028
	Empirical SIMEX	0.013	0.003	-0.0004
	Generalized Regression Calibration	0.018	0.010	0.002
	Weighted Regression Calibration	0.020	0.007	0.005
	Generalized SIMEX	0.023	-0.047	-0.055
	Generalized SIMEX (Combined Proxies)	0.019	0.005	0.001
X1	Standard Regression Calibration	0.003	0.025	0.025
	Standard SIMEX	0.718	0.176	0.105
	Empirical SIMEX	0.003	0.0001	-0.002
	Generalized Regression Calibration	0.003	0.002	0.003
	Weighted Regression Calibration	0.004	0.003	0.002
	Generalized SIMEX	0.005	0.006	0.002
	Generalized SIMEX (Combined Proxies)	0.004	0.003	0.003
X2	Standard Regression Calibration	0.032	-0.175	-0.173
	Standard SIMEX	0.417	-0.630	-0.608
	Empirical SIMEX	0.001	-0.001	-0.001
	Generalized Regression Calibration	0.002	-0.003	-0.002
	Weighted Regression Calibration	0.002	-0.002	-0.002
	Generalized SIMEX	0.002	0.001	0.003
	Generalized SIMEX (Combined Proxies)	0.002	-0.002	-0.0012
X3	Standard Regression Calibration	0.002	-0.037	-0.037
	Standard SIMEX	0.169	-0.261	-0.206
	Empirical SIMEX	0.001	-0.001	-0.001
	Generalized Regression Calibration	0.001	-0.001	-0.001
	Weighted Regression Calibration	0.001	-0.002	-0.003
	Generalized SIMEX	0.001	-0.004	-0.003
	Generalized SIMEX (Combined Proxies)	0.001	-0.002	-0.001

$N(0,1)$, where the errors are all independent. For X_3^* we selected 50% of the observations to be missing.

In this setting we compare standard regression calibration, standard SIMEX, and the empirical SIMEX, with the best performing generalized techniques from the previous simulation (generalized regression calibration with fixed weights, and generalized SIMEX with combined proxies). We include results for both the generalized regression calibration where Z was used as informative, and where this relationship was ignored in estimating the correction parameters. The results regarding the MSE and bias of these estimators are summarized in Table 3.2.

Table 3.2: Results from a simulation of a gamma, log-linear regression model comparing the generalized and standard techniques. The table show the MSE, mean bias, and median bias for the intercept and slope parameters that were estimated in 1000 ($n = 5000$) replicated simulations.

			Bias	
	Method	MSE	Mean	Median
Intercept	Standard Regression Calibration	0.106	-0.325	-0.324
	Standard SIMEX	0.004	0.047	0.047
	Empirical SIMEX	0.026	0.152	0.146
	Generalized Regression Calibration	0.105	-0.323	-0.323
	Generalized Regression Calibration (no Z)	0.103	-0.319	-0.320
	Generalized SIMEX (Combined Proxies)	0.002	0.003	0.001
X	Standard Regression Calibration	0.005	-0.055	-0.053
	Standard SIMEX	0.013	-0.094	-0.093
	Empirical SIMEX	0.614	-0.756	-0.726
	Generalized Regression Calibration	0.002	-0.003	-0.001
	Generalized Regression Calibration (no Z)	0.002	-0.002	-0.001
	Generalized SIMEX (Combined Proxies)	0.004	-0.004	-0.004
Z	Standard Regression Calibration	0.003	0.001	0.002
	Standard SIMEX	0.286	0.023	0.004
	Empirical SIMEX	3.95	0.038	0.006
	Generalized Regression Calibration	0.004	0.001	-0.002
	Generalized Regression Calibration (no Z)	0.003	-0.012	-0.013
	Generalized SIMEX (Combined Proxies)	12.2	0.116	0.001

We note first that for the intercept the methods based on SIMEX perform substantially better, which is to be expected based on the consistency theory. The generalized regression

calibration procedures perform well for the slope parameters, and we see relatively similar results between the estimators which use Z, and those which do not. In this setting we see, on average, an improved MSE and bias with the generalized corrections over the corresponding standard corrections.

It is also worth drawing attention to the fact that, in this setting, the empirical SIMEX sees a dramatic reduction in its performance relative to the other techniques. This can be explained through the lack of normally distributed errors, which the other techniques are more resilient to. Finally, we draw attention to the MSE of the generalized SIMEX for the Z slope parameter, particularly in comparison to the median bias. This large MSE is being driven by several simulation runs which are extreme outliers, and which are not particularly indicative of an actual application of this method. To explain this note that we used the same extrapolant for each iteration of the simulation, without checking the fit (as this would require 1000 different selections for these simulations). However, investigating the outliers, it is clear in these simulation runs the extrapolant is over-fitting noise. An analyst conducting such an analysis would be unlikely to see this degraded performance, as they would be assessing the extrapolant fit directly.

3.8.3 Logistic Regression Models

Finally, we consider a logistic regression model. We take the true covariate $X \sim N(3, 1)$, with $Y \sim \text{Binom}(\text{expit}(0.5 - 0.5X))$. We generate three proxies where $X_1^*, X_2^* \overset{iid}{\sim} X + N(0, 1)$, and $X_3^* = 0.5 + 0.5X + U_3$ where $U_3 \sim \text{Unif}(-0.5, 0.5)$. We select 80% of X_2^* to be missing. In these simulations we compare the results of the generalized estimators, using either all of (X_1^*, X_2^*, X_3^*) or only the iid replicates (X_1^*, X_2^*) for the corrections (labelled "All" and "IID" respectively). Further, we continue to differentiate between the weighted generalized regression calibration and the standard version, as well as the SIMEX estimator that averages proxies ("Combined Proxies") versus the one which averages estimates. In Table 3.1 we observe the MSE and the bias of both the parameter estimates and in Figure 3.1 we observe the estimated probabilities.

The results demonstrate the bias reduction and effective probability estimates of both techniques in logistic regression. Moreover, these simulations demonstrate how biased (using η_0 and η_1) proxies can stabilize estimators. We note that for almost all estimators, the estimators which used all of the proxies (despite the bias) resulted in a reduced MSE compared to the corresponding correction relying on only the iid replicates. There does appear to be evidence of trading off bias and variance within these estimators. The biased replicates appear to introduce slightly larger bias in the estimators, on average, which is

Table 3.3: The estimated parameter values for the intercept and the slope across the different methods. The true values are indicated using a dotted line. Outliers are displayed as filled in circles. Note that the X axes are different for each set of box plots.

			Bias	
	Method	MSE	Mean	Median
Intercept	Generalized Regression Calibration (All)	0.018	0.005	0.010
	Generalized Regression Calibration (IID)	0.025	0.004	0.009
	Weighted Regression Calibration (All)	0.076	0.121	0.131
	Weighted Regression Calibration (IID)	0.025	0.004	0.010
	Generalized SIMEX (All)	0.032	-0.011	0.004
	Generalized SIMEX (IID)	0.059	-0.025	-0.005
	Generalized SIMEX (Combined Proxies, All)	0.020	0.032	0.038
	Generalized SIMEX (Combined Proxies, Reps)	0.033	-0.012	0.002
X	Generalized Regression Calibration (All)	0.002	-0.007	-0.008
	Generalized Regression Calibration (IID)	0.003	-0.010	-0.011
	Weighted Regression Calibration (All)	0.009	-0.049	-0.053
	Weighted Regression Calibration (IID)	0.003	-0.010	-0.011
	Generalized SIMEX (All)	0.004	0.005	-0.001
	Generalized SIMEX (IID)	0.007	0.011	0.003
	Generalized SIMEX (Combined Proxies, All)	0.002	-0.011	-0.013
	Generalized SIMEX (Combined Proxies, Reps)	0.004	0.005	0.001

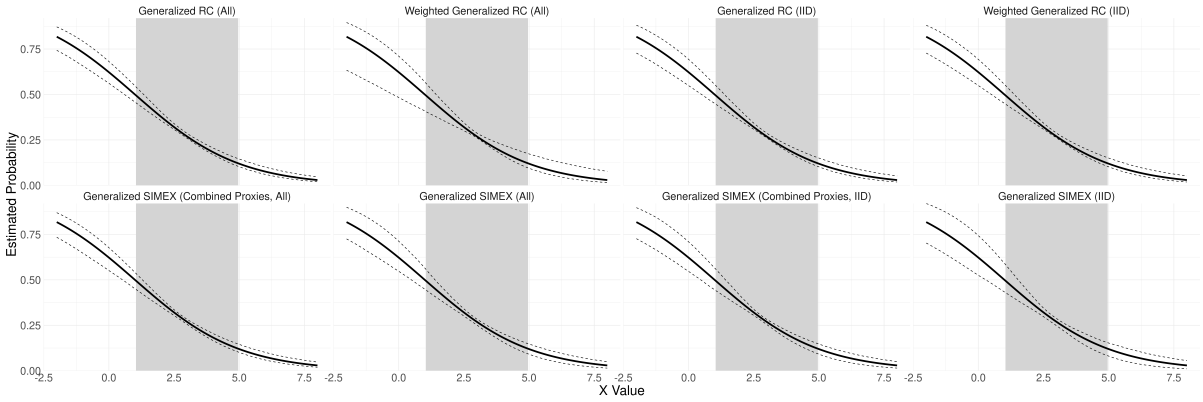

Figure 3.1: The estimated 95% prediction interval for the estimated probabilities (given by the dotted lines) around the true probabilities (given by the solid line), across various values for X. The shaded regions indicate the 95% central values for X, indicating the most likely values for the covariate to take.

made-up for by the reduced variance. We see that the estimated probabilities tend to be correct across any of the methods, with slightly reduced interval widths when making use of the complete data, rather than only the replicated measurements.

3.9 Extensions to Other Methodologies

As introduced in Chapter 2, moment reconstruction is a plug-in technique, similar in spirit to regression calibration, which requires case-specific derivations. We present the results of moment reconstruction in a logistic regression, a case where the moment reconstruction estimators are consistent and the regression calibration corrections are not. The primary motivation for this presentation is demonstrating how the identification of parameters as in Section 3.5, and the related results, can be extended to exact correction methods.

Assume that $X|Y = y \sim N(\mu + y\Delta, \Sigma_{XX})$. Then, for each observation, moment reconstruction forms

$$\widehat{X}_{\mathrm{MR}}(X^*, Y) = \frac{1}{\eta_{1\cdot}} \left\{ \left(E[X^*|Y](I - \widetilde{G}(Y)) + X^* \widetilde{G}(Y) \right) - \eta_{0\cdot} \right\},$$

where $\widetilde{G}(Y) = \eta_{1\cdot} G(Y)$, $\eta_{l\cdot} = \sum_{j=1}^{k} \alpha_j \eta_{lj}$ for $l = 1, 2$, and

$$G(Y) = \operatorname{cov}(X^*|Y)^{-1/2} \operatorname{cov}(X|Y)^{1/2}.$$

This results in $\widehat{X}_{MR}$ having the same first two conditional moments (given Y) as X does. This setup readily presents M-estimators, extending the quantities in Section 3.5. By assumption, Σ_{XX} is both the conditional and unconditional variance of X, which means it is estimated in ξ. Further, $\eta_{l\cdot}$ contain only α_j and parameters estimated in ξ. This leaves $\operatorname{cov}(X^*|Y)$ and $E[X^*|Y]$ to be estimated.

To do so, we can form standard joint M-estimators. Take Θ_1, Θ_2, Θ_3, and Θ_4 to be given by $E[X^*|Y = 1]$, $E[X^*|Y = 0]$, $\operatorname{cov}(X^*|Y = 1)$, and $\operatorname{cov}(X^*|Y = 0)$, respectively. Moreover, assume that the α_j are fixed.[3] Then we can take

$$\begin{aligned}
n^{-1} \sum_{i=1}^{n} y_i \sum_{j=1}^{K} \alpha_j X_{ij}^* - \Theta_1 &= 0; \\
n^{-1} \sum_{i=1}^{n} (1 - y_i) \sum_{j=1}^{K} \alpha_j X_{ij}^* - \Theta_2 &= 0; \\
n^{-1} \sum_{i=1}^{n} y_i \left(\sum_{j=1}^{k} \alpha_j X_{ij}^* - \Theta_1 \right)^2 - \Theta_3 &= 0; \\
n^{-1} \sum_{i=1}^{n} (1 - y_i) \left(\sum_{j=1}^{k} \alpha_j X_{ij}^* - \Theta_2 \right)^2 - \Theta_4 &= 0.
\end{aligned}$$

Denoting the probability $P(Y = 1|X = x) = \operatorname{expit}(\beta_0 + \beta_1 x)$, the logistic regression estimators for β_0 and β_1 are $\widehat{\beta}_0$ and $\widehat{\beta}_1$, with these estimators simultaneously solving

$$n^{-1} \sum_{i=1}^{n} y_i - \operatorname{expit}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i \right) = 0,$$

and

$$n^{-1} \sum_{i=1}^{n} (y_i - \operatorname{expit}\left(\widehat{\beta}_0 + \widehat{\alpha}_1 x_i \right)) x_i = 0.$$

[3] These can estimated in a similar way, if need be.

The moment reconstruction procedure replaces x_i in the above estimating equations with

$$\begin{aligned}
&\widehat{x}_{i,\mathrm{MR}} \\
&\quad = \left(\sum_{j=1}^{k} \alpha_j \eta_{1j}\right)^{-1} \left\{ [y_i \Theta_1 + (1-y_i)\Theta_2] \left[1 - \left(\sum_{j=1}^{k} \alpha_j \eta_{1j}\right) \left(\frac{\Sigma_{XX}}{y_i \Theta_3 + (1-y_i)\Theta_4} \right)^{1/2} \right] \right. \\
&\qquad \left. + \left(\sum_{j=1}^{k} \alpha_j x_{ij}^*\right) \left(\sum_{j=1}^{k} \alpha_j \eta_{1j}\right) \left(\frac{\Sigma_{XX}}{y_i \Theta_3 + (1-y_i)\Theta_4} \right)^{1/2} - \sum_{j=1}^{k} \alpha_j \eta_{0j} \right\}.
\end{aligned}$$

This expression can be inserted into the M-estimators for $\widehat{\beta}_\cdot$, and stacked with the previously discussed M-estimators, allowing for the derivation of an asymptotic distribution. Due to normality, the distributions $(\widehat{X}_{\mathrm{MR}}, Y)$ and (X, Y) are equivalent, and so this estimator will be consistent and asymptotically normal for the true parameter values. Further, the solutions to the M-estimators regarding the parameters in $g_i(\cdot)$ and the estimators for $\Theta_\cdot$ are expressible in closed form, and are functionally independent of $\beta_\cdot$. As a result, they can be solved for first and used to compute $\widehat{x}_{i,\mathrm{MR}}$, before performing a logistic regression.

In order to implement this in practice, or compute closed form expressions for the asymptotic standard errors, we need to make concrete assumptions regarding the repeated measurements that are available and the values of $\{\alpha_j\}$. These data must conform to the identifiability conditions making $g_i(\cdot)$ computable. With a fixed data structure, we can apply Lemma 3.5.2 for the asymptotic distribution.

3.10 Data Analysis

We now apply the generalized methods to data from the Framingham Heart Study. Our analysis is motivated by Carroll, Ruppert, Stefanski, and Crainiceanu [7], where the authors use a logistic regression model to estimate the impact of, age, smoking status, serum cholesterol, and long-term SBP on the likelihood of developing CHD. Our analysis follows a different subset from the FHS, which is made available as a teaching dataset, by the NHLBI [62]. Our subset is not restricted to male participants, and so we use sex as an explanatory factor as well.

Our analysis follows 2876 individuals, aged 32–69, across three separate examinations. We take the patients' sex, age, and smoking status to be error-free, and assume that the serum cholesterol levels and systolic blood pressure are prone to error. Following Cornfield [17] and Carroll et al. [9] we transform the blood pressure measurements to be in-

cluded in the model as $\log(\text{SBP} - 50)$ and the cholesterol measurements to be included as $\log(\text{Cholesterol})$. These data are subject to incomplete replication. Of the 2876 total participants, systolic blood pressure measurements were available for all patients at the first visit, but missing for 153, and 390 patients at visits two, and three respectively. For cholesterol, at visits one, two, and three, there are 26, 256, and 538 patients without repeated measurements, respectively. Considering only those with the repeated measurements taken, at the first visit the mean (transformed) SBP was 4.329 and the mean (transformed) cholesterol was 5.437, with observed variances of 0.052 and 0.033, respectively. This is compared to means (variances) of SBP and cholesterol at the second visit of 4.389 (0.054) and 5.503 (0.030), and at the third visit 4.440 (0.057) and 5.456 (0.033), respectively.

To assess the validity of the regression calibration methodology, we consider plots of the various proxies against one another. For the SIMEX correction, we plot $\widehat{\Theta}(\lambda)$ versus λ to choose the extrapolants. These diagnostic plots are presented in Figures 3.2 and 3.3, respectively. We can see that there is an approximately linear relationship between the various proxies which suggests the use of a linear calibration function is appropriate. This is further emphasized in the discussion by Carroll, Ruppert, Stefanski, and Crainiceanu [7], where it is noted that the transformed blood pressure covariates appear approximately normal. In the SIMEX diagnostic plots, across all settings, there appears to be a quadratic relationship between the estimated slope parameters and λ.

We compare several analyses, all of which use the main effects model in a standard logistic regression. We consider a naive analysis, which takes the mean response from the visits for both cholesterol and blood pressure as the explanatory factors, a standard regression calibration analysis which implicitly assumes that the repeated measurements are iid, and several scenarios for the generalized procedures presented. We use different assumptions for J_0, the proxies which have $\eta_{0j} = 0$, and J_1, those with $\eta_{1j} = 1$. For regression calibration we consider four scenarios, two with $J_0 = \{1, 2, 3\}$, where $J_1 = \{1, 2, 3\}$ or $J_1 = \{1, 2\}$, in addition to two with $J_0 = \{2\}$, with $J_1 = \{1, 3\}$ or $J_1 = \{2, 3\}$. We conduct two SIMEX analyses, one with $J_0 = J_1 = \{1, 2, 3\}$, and one with $J_0 = \{2\}$ and $J_1 = \{1, 3\}$. The SIMEX procedures are restricted in their consideration due, in part, to the concerns regarding the validity of M_j as a variance matrix. Many plausible settings lead to singular matrices as estimates for M_j, which in turn rules out the use of the modified SIMEX under those assumptions. The SIMEX procedures used a quadratic extrapolant for both the SBP and the cholesterol terms.

The results of these analyses are displayed in Table 3.4, where the slope parameter estimates for the transformed systolic blood pressure and the transformed cholesterol are presented, along with 95% bootstrapped confidence intervals. The bootstrap confidence intervals are derived from 1000 bootstrap replicates in each scenario. Across the various

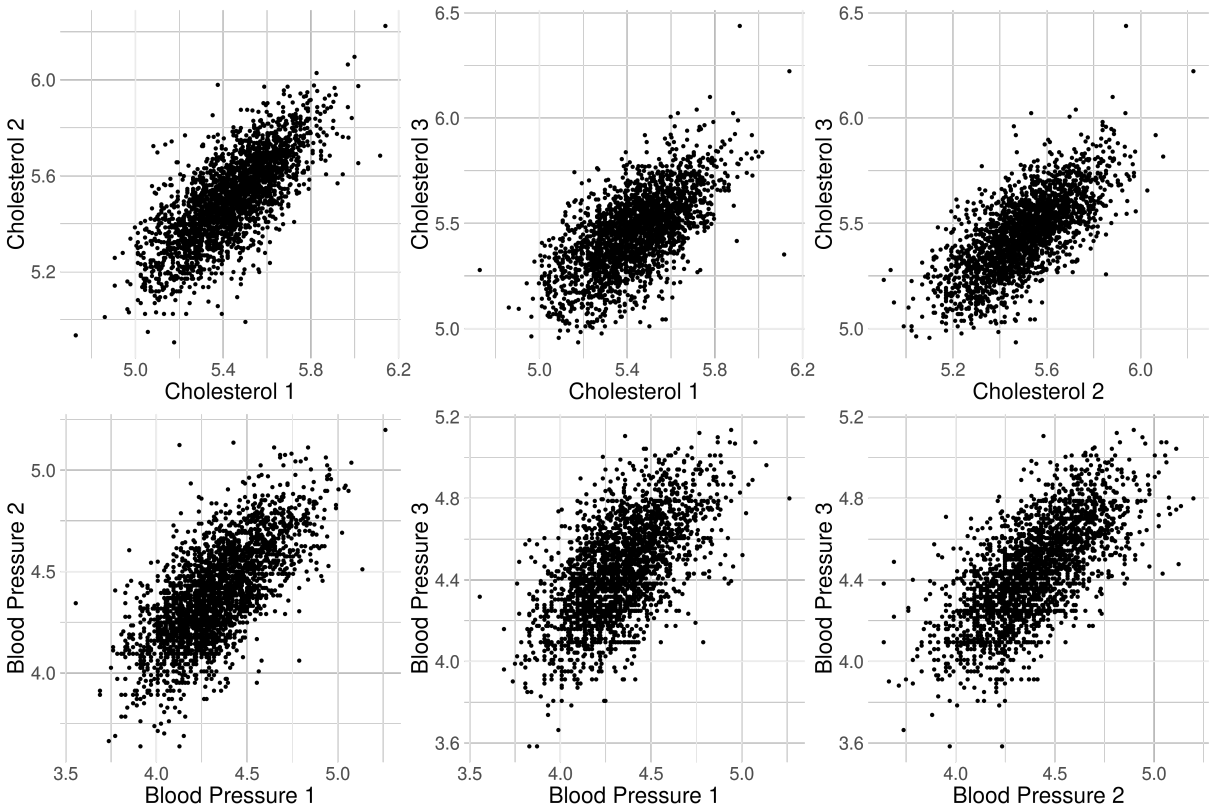

Figure 3.2: Plots showcasing the approximate linearity between the three proxy measurements for cholesterol (top row) and the three proxy measurements for blood pressure (bottom row). The apparent linearity in these plots gives evidence for the fact that $E[X|X^*]$ is well approximated by a linear relationship.

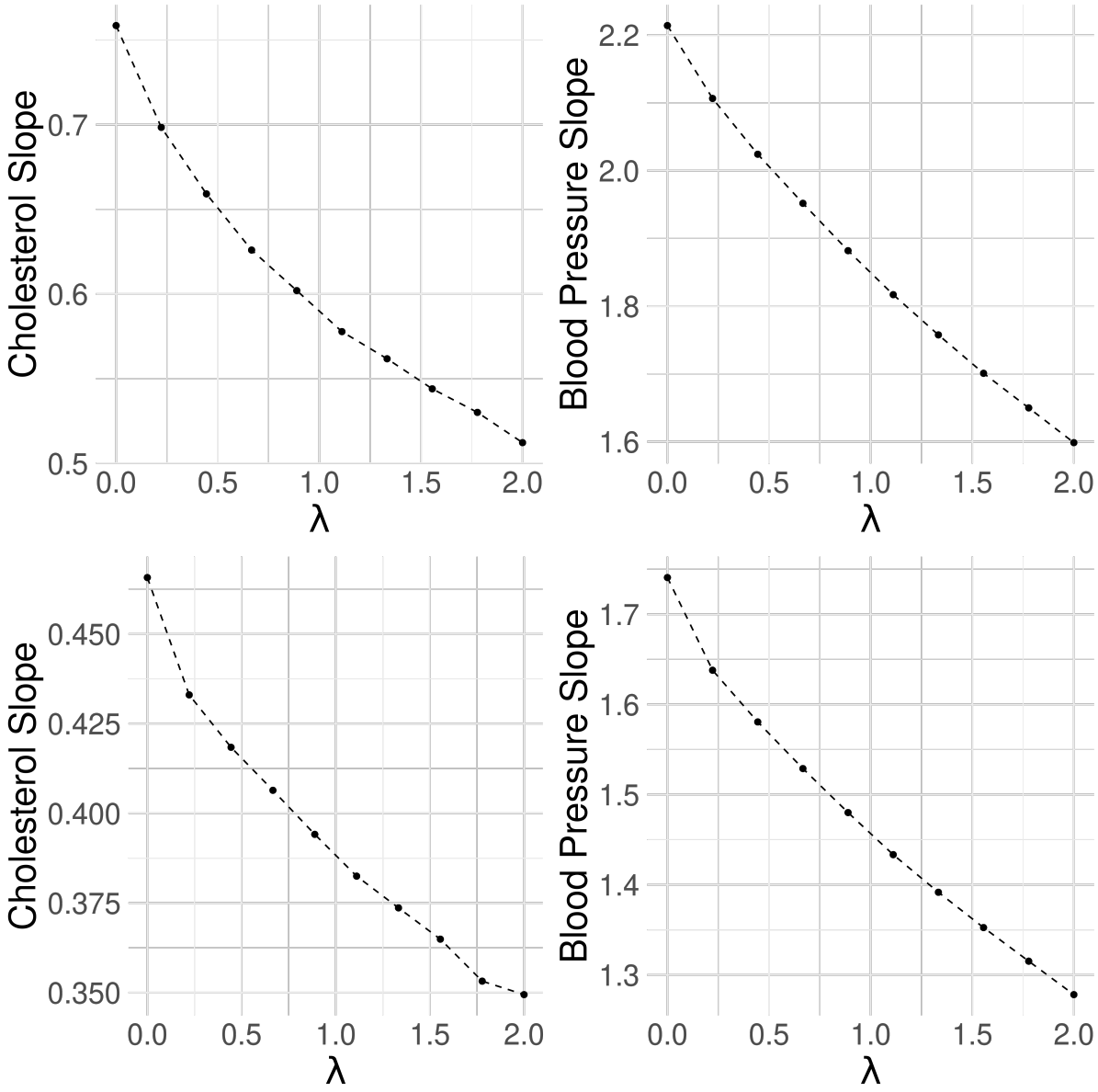

Figure 3.3: Plots of $\widehat{\Theta}(\lambda)$ against λ, over the fitted range of the data, for the slopes on both the cholesterol and blood pressure coefficients. The top row contains the results assuming $J_0 = J_1 = \{1, 3\}$ while the bottom row contains those assuming $J_0 = 2$ and $J_1 = \{1, 3\}$. These plots illustrate the approximate extrapolant shape to be used in the SIMEX procedure. The results suggest that a quadratic extrapolant may be effective.

different sets of assumptions, we observe some variability in the point estimates for both factors, with more substantial variability in the cholesterol measurements. While none of the methods find the effect of cholesterol to be significant at a 95% confidence level, the implied level of significance varies across the scenarios.

Table 3.4: Estimated slope parameter for the SBP and cholesterol terms, in the FHS, comparing the generalized regression calibration and SIMEX methodologies to a naive analysis and standard regression calibration. The point estimates and 95% confidence interval are shown, where the intervals are based on a bias corrected bootstrap procedure with 1000 bootstrap replicates.

Method	Blood Pressure	Cholesterol
Naive	2.250 (1.696, 2.837)	0.670 (-0.083, 1.575)
Standard Regression Calibration	2.811 (2.104, 3.591)	0.753 (-0.177, 1.866)
Generalized Regression Calibration		
$J_0 = J_1 = \{1, 2, 3\}$	2.688 (2.005, 3.417)	0.723 (-0.171, 1.790)
$J_0 = \{1, 2, 3\}$; $J_1 = \{1, 2\}$	2.673 (1.992, 3.412)	0.935 (-0.138, 2.207)
$J_0 = 2$; $J_1 = \{1, 3\}$	2.635 (1.917, 3.415)	0.732 (-0.168, 1.808)
$J_0 = 2$; $J_1 = \{2, 3\}$	2.785 (2.097, 3.550)	0.347 (-0.200, 1.276)
Generalized Simulation Extrapolation		
$J_0 = J_1 = \{1, 2, 3\}$	2.674 (1.626, 5.892)	1.000 (-0.476, 15.182)
$J_0 = 2$; $J_1 = \{1, 3\}$	2.096 (1.168, 6.406)	0.567 (-2.454, 3.764)

Chapter 4

Simulation Extrapolation

4.1 Motivation for the Nonparametric SIMEX

A common assumption in the measurement error literature is that errors are normally distributed. This assumption is analytically attractive and holds approximately in many scenarios. However, it is often violated in practice. Bailey [1] investigated the distributions of measurement variability across several fields, including medicine, nuclear physics, and toxicology, and found that differences between measurements of the same quantity are consistent with heavy-tailed t-distributions more so than Gaussian distribution. Further examples of nonnormal error distributions are readily available. McKenzie et al. [55] consider the use of a bivariate Laplace distribution to model the location error present when using GPS collars to study animal movement and habitat use. Bollinger [4] determine that the assumption of normality is strongly violated when looking at the errors in reported annual incomes within the Current Population Survey. Purdom and Holmes [65] use an asymmetric Laplace distribution to model the error distribution in microarray data. Rajan and Desai [66] argue that a t-distribution with two degrees of freedom is the best parametric fit to the error distribution for measurements of galactic rotation speed. Xu, Kim, and Li [99] demonstrate that errors in reported BMI in the Korean Longitudinal Study of Ageing are inconsistent with a normal distribution, exhibiting heavy-tailed behaviour. In nutritional epidemiology, it is often suggested to transform reported intakes in such a way so as to approximate normality of the errors, though often the suggested transformation fails to achieve suitably normal errors [19, 64].

In their work on transformations of non-additive models, Eckert, Carroll, and Wang [24] discuss how transformations can be used to induce an error structure with error terms

that approximate a particular distribution. The argument for such a transformation is that many existing techniques, in addition to assuming an additive error structure further assume normality of the underlying errors. Owing to the fact that normality of errors has been empirically shown to be violated across many domains, understanding the impact of violations of normality on measurement error correction techniques is important. Moreover, developing accessible techniques which can account for non-normal errors is required in order to facilitate the use of valid measurement error corrections in the applied literature.

The understanding of the impact of distributional assumptions is particularly important for techniques which are approximately consistent. This is because approximately consistent techniques are often presented as being "better than nothing". If distributional assumptions are required for this to hold, then these assumptions should be clearly communicated and, ideally, easy to test. When these assumptions are required, an understanding of how the methods breakdown when they are violated can lend additional confidence to the application of an existing method. In Chapter 3 we discussed the distributional requirements for applying regression calibration to result in consistent estimators. For the BLUP to be valid, we require the conditional mean of X given $\{X^*, Z\}$ to be linear. Conditional normality suffices for this requirement to be met.

This same type of analysis has not previously been conducted for simulation extrapolation. The initial proposal of SIMEX required normally distributed errors [16]. If the errors are not normally distributed, then the pseudo-random error terms will not be absorbed into a unified, normally distributed error term, and the presented simplifications do not occur. The importance of normality was emphasized by Stefanski and Cook [87], and then again by Koul and Song [50], when illustrating the conditions under which SIMEX can produce consistent estimators. In simulations, Yi and He [102] further illustrate the requirement of normality for SIMEX. These results are particularly concerning since the SIMEX estimator not only became inconsistent, but actually performed worse than the naive estimator. Taken together, these theoretical and simulated results seem to establish the need for normality to apply SIMEX.

If it is the case that SIMEX, as it is commonly presented, requires normally distributed errors to serve as a useful correction, it is worth quantifying this as best as we can. Would it be possible to theoretically identify the types of scenarios presented by Yi and He [102], where SIMEX amplified the bias? The work by Koul and Song [50] suggests that the underlying concept behind SIMEX, when modified, can be used to accommodate errors which are not normally distributed. Expanding their work beyond parametric assumptions presents an opportunity for SIMEX to be applied in a wide variety of settings, without the need to test distributional assumptions, or further adapt existing methods.

4.2 Summary of the Proposed Methods

In this chapter we consider a nonparametric extension to the simulation extrapolation method, which we call the NP-SIMEX. The NP-SIMEX functions via the familiar stages also present in the standard parametric SIMEX, where we first simulate the impact of additional measurement error on the estimator of interest, and then we extrapolate to the setting where this error is removed. As the name suggests, the NP-SIMEX uses nonparametric techniques in the simulation step, rendering it resilient to normality violations.

Where the standard SIMEX proceeds by drawing realizations from a standard normal distribution, and then multiplying by $\lambda^{1/2}\sigma_U$ for each value of λ in a grid, the NP-SIMEX makes this nonparametric. Specifically, λ (for integer valued λ) realizations are drawn from the empirical error distribution, for each individual, and are added to the variate of interest. This is taken to be the re-measured version of the surrogate, and we use this to compute the estimator of interest. The remaining procedure is exactly the same: these re-measured variates are used to estimate the parameter of interest across a grid of λ, repeated several times and averaged to reduce uncertainty, and then a parametric extrapolant is fit based on λ. The true estimate is computed by taking $\lambda = -1$ in fitted parametric model.

The primary additional consideration required for the implementation of the NP-SIMEX, as compared to the standard SIMEX, is how we estimate the empirical error distribution. We demonstrate how this can be done with a validation sample (Section 4.5.3), which makes use of the fact that within validation data we observe the errors directly, as well as with replicate measurements (Section 4.5.4) where we need to impose symmetry assumptions on the distribution of errors. Before presenting the nonparametric SIMEX, we first take a deep theoretical look at the need for normality within the standard SIMEX procedure itself, which provides a means of motivating the development of the NP-SIMEX.

4.3 Reframing SIMEX with Characteristic Functions

The importance of normality for SIMEX theoretically, as emphasized by Stefanski and Cook [87] and Koul and Song [50], can be summarized concisely with the following theorem.

Theorem 4.3.1 (Complex Moments are 0)**.** *If U_1 and U_2 are iid, symmetric, absolutely continuous random variables with a finite moment generating function, then $E[(U_1 + \sqrt{-1}U_2)^n] = 0$ for all $n = 1, 2, \ldots$ if and only if U_1 and U_2 are normally distributed with mean 0. (Theorem 3.1 [50])*

The utility of this theorem is that, through the use of a Taylor expansion, it allows for us to show the necessity of normality to render SIMEX operational. If we consider $\Theta = f(X)$, for some "sufficiently smooth" function, f,[1], then we can consider the power series representation of $f(X+U+i\sigma\epsilon)$, where $\sigma^2 = \text{var}(U)$, and $\epsilon \perp \{U, X\}$, with $E[\epsilon] = 0$ and $\text{var}(\epsilon) = 1$. If we take the conditional expectation, given X, then the power series representation (around X) gives

$$E\left[f(X+U+\sqrt{-1}\sigma\epsilon)\middle| X\right] = f(X) + \sum_{j=1}^{\infty} \frac{f^{(j)}(X)\sigma^j}{j!} E\left[(U+\sqrt{-1}\sigma\epsilon)^j\middle| X\right]. \tag{4.3.1}$$

Under correct extrapolant specification, the SIMEX estimator will consistently estimate $E[f(X+U+\sqrt{-1}\sigma\epsilon)|X]$. As a result, applying Theorem 4.3.1 tells us that, if $\epsilon \sim N(0,1)$, then the only way that this estimator is consistent for all such f is if $U \sim N(0,\sigma^2)$. This result applies to the class of "all sufficiently smooth f." In practice, we need not worry about the class of all functions f, simply because a small number of estimators are likely to be the ones under consideration. While it may be the case that normality is required to be able to consider arbitrary functional relationships between Θ and X, we are not typically considering arbitrary functional relationships, and may not require normality.

In extending SIMEX for errors that follow a Laplace distribution, Koul and Song [50] show that, if the characteristic function of $U + \sigma\epsilon$ (denoted $\varphi_{U+\sigma\epsilon}(t)$) tends to 1 as some distributional parameter of ϵ tends towards a constant, then SIMEX estimators will be consistent. In the standard SIMEX, we would take $\epsilon \sim N(0,\lambda)$. We then look at $\varphi_{U+\sigma\epsilon}(t)$ as $\lambda \to -1$. The authors use this argument to show that, by changing the distribution of ϵ, similar strategies can be derived based on an assumed error distribution for U.

If we denote $X_b^*(\lambda) = X^* + \epsilon_\lambda = X + U + \epsilon$, where here λ now specifies a (controllable) parameter of ϵ, then we can consider the characteristic function of $U_\lambda := U + \epsilon_\lambda$, defined as $\varphi_{U_\lambda}(t)$. We are interested in considering this quantity when $\epsilon_\lambda \sim N(0, \lambda\sigma^2)$, where $\text{var}(U) = \sigma^2$. Denoting the joint distribution of $\{Y, X\}$ as $F_{Y,X}$, then by viewing the estimator as a functional on distributions, there exists some $\mathbf{T}$ such that $\Theta = \mathbf{T}(F_{Y,X})$. Moreover, the SIMEX estimator consistently estimates $\Theta_{\text{SIMEX}} = \lim_{\lambda\to-1} \mathbf{T}(F_{Y,X} * F_{0,U_\lambda})$, where $*$ is the convolution operator.

If $\mathbf{T}(\cdot)$ is such that $\lim_{\lambda\to-1} \mathbf{T}(F_{Y,X} * F_{0,U_\lambda}) = \mathbf{T}\left(\lim_{\lambda\to-1} F_{Y,X} * F_{0,U_\lambda}\right)$, then our consideration of the characteristic function of U_λ becomes entirely natural. A function over

[1]As discussed in Stefanski and Cook [87] the restrictions on f are stronger than normal regularity conditions here. We can, for instance, take f to be analytic on the real line, and have a power series in which expectation and summation can be interchanged.

distributions can be represented as a function on the characteristic functions themselves. The characteristic function of a convolution (assuming independence), is given by the product of characteristic functions. This formulation then demonstrates why having $\varphi_{U_\lambda}(t) \to 1$ is important for the consistency of SIMEX. This framework can be used to more deeply explore the theoretical behaviour of SIMEX estimators broadly.

4.4 Asymptotic Analysis of the Standard SIMEX

4.4.1 Approximations of the Characteristic Function

In Table 4.1, we show the characteristic function of U_λ and of $U^* = \lim_{\lambda \to -1} U_\lambda$, based on various distributions of U. Under the assumption of normality, the characteristic function equals 1 exactly in the limit. In all other settings it does not. The limiting characteristic functions are not generally valid characteristic functions, though many are close approximations to 1 around $t = 0$. To see this concretely, we take Taylor approximations to these functions, in a neighbourhood of $t = 0$. The results are presented in Table 4.2.

Table 4.1: Limits of the characteristic function, as $\lambda \to -1$ for the convolution of the errors and pseudo-errors, assuming a normally distributed pseudo-error with different error distributions. Error distributions are parameterized such that $E[U] = 0$ and $\text{var}(U) = \sigma^2$.

Distribution	φ_{U_λ}	φ_{U^*}
Normal	$\exp\left(-\frac{1}{2}\sigma^2 t^2(1+\lambda)\right)$	1
Laplace	$\left(1+\frac{\sigma^2 t^2}{2}\right)^{-1} \exp\left(-\frac{\lambda\sigma^2 t^2}{2}\right)$	$\left(1+\frac{\sigma^2 t^2}{2}\right)^{-1} \exp\left(\frac{\sigma^2 t^2}{2}\right)$
Uniform	$\left(\frac{\exp(it\sqrt{3}\sigma)-\exp(-it\sqrt{3}\sigma)}{2it\sqrt{3}\sigma}\right) \exp\left(-\frac{\lambda\sigma^2 t^2}{2}\right)$	$\frac{\left(e^{it\sqrt{3}\sigma}-e^{-it\sqrt{3}\sigma}\right)\exp\left(\frac{\sigma^2 t^2}{2}\right)}{\left(it2\sqrt{3}\sigma\right)}$
Discrete Unif.	$\frac{1}{2}\left(e^{it\sigma}+e^{-it\sigma}\right)\exp\left(-\frac{\lambda\sigma^2 t^2}{2}\right)$	$\frac{1}{2}\left(e^{it\sigma}+e^{-it\sigma}\right)\exp\left(\frac{\sigma^2 t^2}{2}\right)$
Exponential	$(1-it\sigma)^{-1}\exp\left(\frac{t\sigma}{2}\left(-\lambda t\sigma-2i\right)\right)$	$(1-it\sigma)^{-1}\exp\left(\frac{t\sigma}{2}\left(t\sigma-2i\right)\right)$

These approximations show that, under many error distributions, the limiting characteristic function approximates 1 fairly closely. If we consider the earlier Taylor expansion

(Equation 4.3.1) then, under most of the presented error models, we will see that

$$E[f(X+U^*)|X] \approx f(X) + \frac{f^{(4)}(X)\sigma^4}{4!}E[U_\lambda^4|X] \approx f(X) + f^{(4)}(X)\sigma^8,$$

where the approximation is due to the fact that $E[U_\lambda^4|X] \approx c\sigma^4$ for some constant c (dependent on the error distribution), and is valid to a term that is $O(\sigma^{10})$. This approximation will often be adequate, especially when $\sigma^2 \approx 0$.

Table 4.2: Approximations of the characteristic function for the convolution of the errors and pseudo-errors, assuming a normally distributed pseudo-error with different error distributions. Error distributions are parameterized such that $E[U] = 0$ and $\text{var}(U) = \sigma^2$.

Distribution	Approximation to $\varphi_{U_\lambda}(t)$	Approximation to $\varphi_{U^*}(t)$
Normal	$1 - \frac{\sigma^2t^2}{2}(\lambda+1) + \frac{\sigma^4t^4}{8}(\lambda+1)^2 + O(\sigma^6t^6)$	1
Laplace	$1 - \frac{\sigma^2t^2}{2}(\lambda+1) + \frac{\sigma^4t^4}{40}(5\lambda^2+10\lambda+3) + O(\sigma^6t^6)$	$1 + \frac{1}{8}\sigma^4t^4 + O\left(\sigma^6t^6\right)$
Uniform	$1 - \frac{\sigma^2t^2}{2}(\lambda+1) + \frac{\sigma^4t^4}{8}(\lambda^2+2\lambda+2) + O(\sigma^6t^6)$	$1 - \frac{1}{20}\sigma^4t^4 + O(\sigma^6t^6)$
Discrete Unif.	$1 - \frac{\sigma^2t^2}{2}(\lambda+1) + \frac{\sigma^4t^4}{24}(3\lambda^2+6\lambda+1) + O(\sigma^6t^6)$	$1 - \frac{1}{12}\sigma^4t^4 + O(\sigma^6t^6)$
Exponential	$1 - \frac{\sigma^2t^2}{2}(\lambda+1) - \frac{i\sigma^3t^3}{3} + O(\sigma^4t^4)$	$1 - \frac{i}{3}\sigma^3t^3 + O(\sigma^4t^4)$

An alternative method to assess the quality of this approximation is through the consideration of moments of $X + U_\lambda$. To describe the dependence of this approximation on the functional $\mathbf{T}$, we focus on the class that $\mathbf{T}$ belongs to. Define a space of distributional functions, $\mathscr{T}_m$, such that for every $\mathbf{T} \in \mathscr{T}_m$, only the first m moments of the distribution are relevant. These are quantities which can be consistently estimated using sample moments of order 1 through m. The m-th moment of a random variable, if it exists, is given by $E[Z^m] = \sqrt{-1}^{-n}\varphi_Z^{(m)}(0)$. From the Taylor series approximation for the limiting characteristic functions of U^*, it is clear that all moments m up to the second included term (in Table 4.2) are 0. Up to this term, we have

$$\varphi_{X+U^*}^{(m)}(0) = \sum_{j=0}^{m}\binom{m}{j}\varphi_X^{(m)}(0)\varphi_{U^*}^{(m-j)}(0) = \varphi_X^{(m)}(0),$$

and as a result we will have $E[(X+U^*)^m] \approx E[X^m]$. If we call the order of this second included term M, then, any functions $\mathbf{T} \in \mathscr{T}_{M-1}$ should have no error in the limiting term. For $\mathbf{T} \in \mathscr{T}_M$, the limiting term will be biased, since $\varphi^{(M)}_{X+U^*}(0) = \varphi_X(0) + c\sigma^M + o(\sigma^M)$.

4.4.2 Demonstration of Excess Bias

To motivate their Laplace modified SIMEX, Koul and Song [50] used an estimator of the fourth moment.[2] They took $\widehat{\mu}_4 = n^{-1}\sum_{i=1}^n X_i$, and used normal pseudo-errors with an underlying Laplace distribution. We can view this directly as $f(X)$, where we note that $f^{(k)}(X) = 0$ for all $k > 4$. As a result, the bias is going to be exactly the $f^{(4)}(X)$ term, in the Taylor expansion. Consulting Table 4.2, and following our previous argument, we get that $f^{(4)}(X) = 24$, $E[U^4_\lambda|X] = 3\sigma^4$, so that the bias will be $3\sigma^4$. Instead, we can view this as $\mathbf{T} \in \mathscr{T}_4$. Here we note that $\mathbf{T}$ is linear in the 4-th moment, and as a result the bias is going to be given simply as the bias in the fourth moment. Table 4.2 gives that this will be $3\sigma^4$. In motivating their modified method, the authors work through the algebra to arrive at the conclusion that the exact bias of this estimator is $3\sigma^4$, as this theory predicts.

4.4.3 Decomposition of the Asymptotic Bias

Until now we have assumed that $\mathbf{T}$ is known and correctly specified. We cannot (in general) compute $\mathbf{T}$ as a function of λ in a closed form, and instead specify it according to an assumed parametric form $\mathcal{G}$. We fit $\widehat{\Theta}_{\text{SIMEX}} = \widehat{\mathcal{G}}(-1)$, where some model is posed that extrapolates $\mathbf{T}(F_{Y,X} * F_{0,U_\lambda})$ to a complete curve, allowing us to take $\lambda \to -1$. Determining an exact extrapolant is unlikely to be a straightforward task for most settings. The fact that an exact extrapolant may not be available introduces another source of possible bias.

There are thus two possible sources of asymptotic bias. The first is the bias derived from taking $\lim_{\lambda\to-1}\mathbf{T}(F_{Y,X} * F_{0,U_\lambda})$ as a proxy for $\mathbf{T}(F_{Y,X})$. In the previous sections this bias was shown to be 0 when the error was normal, and we discussed a mechanism for approximating this when errors are non-normal. The second component of the bias comes from using $\widehat{\mathcal{G}}(\lambda)$ as an estimator for $\mathbf{T}(F_{Y,X} * F_{0,U_\lambda})$. This will be a more traditional model misspecification problem, where we are considering asymptotic bias from extrapolation. We take $\text{ABias}(\widehat{\Theta}_{\text{SIMEX}})$ to be,

$$\widehat{\mathcal{G}}(-1) - \mathbf{T}(F_{Y,X}) = \Big[\widehat{\mathcal{G}}(-1) - \mathbf{T}(F_{Y,X} * F_{0,U^*})\Big] + \Big[\mathbf{T}(F_{Y,X} * F_{0,U^*}) - \mathbf{T}(F_{Y,X})\Big].$$

[2] We consider this example in more detail in Section 4.5.

Note that this same decomposition can be used for any λ. Since the naive estimator is equal to the aforementioned estimator when $\lambda = 0$, we get that

$$\text{ABias}(\widehat{\Theta}_{\text{Naive}}) = \left[\widehat{\mathcal{G}}(0) - \mathbf{T}(F_{Y,X} * F_{0,U_0})\right] + \left[\mathbf{T}(F_{Y,X} * F_{0,U_0}) - \mathbf{T}(F_{Y,X})\right].$$

As a general rule if $\lambda \geq 0$, the bias decomposition will have 0 for the first component since we can actually fit the model $\mathbf{T}(F_{Y,X} * F_{0,U_\lambda})$ directly.

In the original proposal of SIMEX, Cook and Stefanski [16] derive the fact that the asymptotic bias of $\widehat{\Theta}_{\text{SIMEX}}$ is of order $O(\sigma^6)$, when using the quadratic or nonlinear extrapolants, and $O(\sigma^4)$, when using a linear extrapolant. Since they assumed normality, this is the order of the first term in the expression for asymptotic bias.

Viewed as a mechanism to reduce the bias present when measurement error is an issue, SIMEX need not produce unbiased or consistent estimators in order to be useful. Any situation where SIMEX produces a substantive decrease in bias compared to the naive estimator, regardless of consistency claims, is a situation where it may be useful.

As an example, consider estimating the fourth moment of X, denoted μ_4, using the fourth empirical moment. Then,

$$E[\widehat{\Theta}_b(\lambda)] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_{bi}^*(\lambda)^4\right] = \mu_4 + 6\mu_2\sigma^2(1+\lambda) + E(U^4) + 6\sigma^4\lambda + 3\sigma^4\lambda^2,$$

which depends on the distribution of U only in its fourth moment. The extrapolant in this case is quadratic in λ. We can also see from this expression the conditions under which the estimator removes bias, asymptotically. If $\lambda = -1$, then the expression simplifies to $\mu_4 + E[U^4] - 3\sigma^4$. The bias will be 0 when $E[U^4] = 3\sigma^4$, meaning that normality will suffice, though it is not necessary.[3] Since the true extrapolant is quadratic, we can consider misspecifying it as a linear extrapolant. Re-writing the above model, we get

$$\mathcal{G}(\lambda) = \underbrace{\mu_4 + 6\mu_2 + E[U^4]}_{:=a_0} + \underbrace{6(\sigma^4 + \mu_2)}_{:=b_0}\lambda + \underbrace{3\sigma^4}_{:=c_0}\lambda^2,$$

which we estimate linearly as $\widehat{\mathcal{G}}(\lambda) = \widehat{\alpha} + \widehat{\beta}\lambda$. We can fit this line using two observations, $\lambda = 0$ and $\lambda = \lambda_1 > 0$. The fitted values will be $\widehat{\alpha} = a_0$ and $\widehat{\beta} = b_0 + c_0\lambda_1$.

The choice of λ_1 will dictate the exact fit. Extrapolating to $\lambda = -1$ gives the estimate

[3] $U \sim \text{Unif}\left\{\pm\sqrt{\sigma^2 + \sqrt{\sigma^2(3-\sigma^2)}}, \pm\sqrt{2\sigma^2 + \sqrt{\sigma^2 + \sqrt{\sigma^2(3-\sigma^2)}}}\right\}$ also will suffice.

$\widehat{\mathcal{G}}(-1) = \widehat{\alpha} - \widehat{\beta} = a_0 - b_0 - c_0\lambda_1$. Comparing this to $\mathcal{G}(-1) = a_0 - b_0 + c_0$, the bias, in general, will be $-c_0(\lambda_1+1) = -3\sigma^4(\lambda_1+1)$. This bias is present due to the misspecification of the extrapolant. The total bias of the estimator is

$$\text{ABias}(\widehat{\Theta}_{\text{SIMEX}}) = E[U^4] - 3\sigma^4(\lambda_1 + 1).$$

The naive estimator will consistently estimate $\mu_4 + 6\mu_2\sigma^2 + E[U^4]$. This will almost surely be different from μ_4. The naive estimator will not have any model misspecification bias, and as a result, our total bias is

$$\text{ABias}(\widehat{\Theta}_{\text{Naive}}) = 6\mu_2\sigma^2 + E[U^4].$$

Comparing these results relies on a specification of the distribution of X and U. When $E[U^4] = 3\sigma^4$, then $|\text{ABias}(\widehat{\Theta}_{\text{Naive}})| < |\text{ABias}(\widehat{\Theta}_{\text{SIMEX}})|$ so long as $\frac{\lambda_1}{2}\sigma^2 > \mu_2$. The example is somewhat contrived in that, fitting a linear model would be very unlikely for an analyst considering the plots directly. However, the illustration is important insofar as the sources of bias can be examined.

While it is the case that, generally, the SIMEX estimator will reduce the asymptotic bias associated with approximating Θ by $\mathcal{G}(\lambda)$, it may not be the case that an overall bias reduction is attained. This example suggests a set of tools for considering the behaviour of the SIMEX estimators in real applications. Yi and He [102], when using SIMEX in a proportional odds model, note that "When the measurement error model induces more misspecification, the performance of the SIMEX method can deteriorate more noticeably. Its point estimate can incur a larger bias than that of the naive analysis [...]."

4.4.4 Considerations for λ

Initially, the justification for taking $\lambda = -1$ in SIMEX was motivated by consideration of the variance. This intuition is explicitly justified through a consideration of the characteristic function of the error term, U_λ.

Lemma 4.4.1 (Best Second Order Approximation)**.** *Taking $\lambda = -1$ provides the best second-order approximation to $\varphi_{U_\lambda}(t) = 1$ in a neighbourhood of $t = 0$.*

This result is, in essence, a re-characterization of the intuitive variance explanation. However, our previous discussion suggests that we ought to be concerned with the quality of the approximation beyond second moments. One natural question is whether there is

a λ that is preferable to -1, to extrapolate to. This result extends the idea of Lemma 4.4.1. Here we say that, so long as U is symmetric around 0, there is a region around 0 where $\lambda = -1$ gives the closest (in terms of squared distance) approximation to 1, for characteristic functions taking the form of $\varphi_{U_\lambda}(t)$.

Theorem 4.4.2 (SIMEX Approximation Uniformly Dominates)**.** *Assume that U has a symmetric (about 0) distribution. Then, there exists some $\epsilon > 0$ such that uniformly on $t \in (-\epsilon, \epsilon)$ we will have that $(1 - \varphi_{U_{-1}}(t))^2 \leq (1 - \varphi_{U_\zeta}(t))^2$, for all $\zeta \neq -1$.*

This result suggests that, at least as long as the underlying process of simulation extrapolation is not altered, the intuitive selection of $\lambda = -1$ cannot be improved upon (in terms of the MSE). This result is not necessarily surprising, as taking $\lambda = -1$ is natural given the development of SIMEX generally. However, this is a useful result to know when SIMEX is being used to reduce the bias in a naive estimator. As has been developed throughout this chapter thus far, even under normality violations, it may often be the case that SIMEX serves a useful tool for lessening the impact of error. With this result, an analyst using SIMEX as to reduce bias need not consider alternative values for λ.

4.4.5 Summary of Characteristic Function Framing

These results centre on the key idea that simulation extrapolation, in the case of normal errors, functions predominantly by generating a sequence of (pseudo) random variables with a characteristic function that tends to 1 as λ tends to -1. By taking this property as the defining relationship for SIMEX estimators, we are able to understand the sources of asymptotic bias that arise when normality violations occur. Moreover, recognizing that characteristic functions are directly tied to the moments of a distribution, we can characterize when the standard SIMEX procedure is capable of eliminating all asymptotic bias based on the underlying estimator. Any estimator which depends on the underlying distribution only through the first m moments can be consistently estimated under any error distribution with a Taylor series approximation of the characteristic function equal to $1+O(\sigma^{m+1})$. Estimators which require further moments of the distribution to be computed will be more severely impacted by deviations from normality.

This formulation also allows for a decomposition of the asymptotic bias in the estimators based on both the accuracy of the extrapolant and the convergence of the characteristic function. This decomposition helps to demonstrate how it can be possible for the naive estimators to exhibit less bias asymptotically than the SIMEX corrected estimators. This can be true even in settings where the SIMEX could be exactly consistent, had the extrapolant been correctly specified.

These considerations together help to describe when the SIMEX methodology can be validly applied, and to give further justification to why it works. This formulation also gives rise to a useful extension of SIMEX allowing for consistency of the underlying corrections, under any error distribution. The key to this nonparametric SIMEX is realizing that, using the observed empirical distribution, it is possible to have $\varphi_{U_\lambda}(t) \to 1$ as $\lambda \to -1$.

4.5 Nonparametric Simulation Extrapolation

In order to ensure clarity, the standard SIMEX, which relies on parametric assumptions for consistency, will be referred to as P-SIMEX (*p*arametric SIMEX) during the remainder of this chapter. This is predominantly to distinguish it from the NP-SIMEX (*nonparametric SIMEX*), which is introduced as a generalization.

The P-SIMEX has been described as a *remeasurement method* [63], emphasizing its similarities to bootstrap procedures. We can view the simulated, additional error as "remeasuring" the error-prone proxy from a distribution with variance $(1+\lambda)\sigma_U^2$. To further emphasize the analogy, the P-SIMEX is analogous to the parametric bootstrap, since this procedure of remeasuring occurs on the basis of a parametric assumption. Just as how bootstrap procedures can be made nonparametric by resampling from the empirical distribution, the P-SIMEX can be made nonparametric by *remeasuring* using the empirical error distribution. This allows for the NP-SIMEX to accommodate a wide range of error models, without making any specific distributional assumptions.

The proposed NP-SIMEX stands in contrast to the methods of Koul and Song [50], who propose a parametric SIMEX based on non-normal distributions. Whether in the case of normally distributed errors, or in the more generally proposed methods, the P-SIMEX can be viewed as analogous to a resampling procedure from an estimated, parametric distribution. The distribution of U_{ij} is assumed to be known, and characterized by a parameter, say $U_{ij} \sim F_{\sigma_U^2}$. Then, remeasurement proceeds by drawing independent realizations from $F_{\widehat{\sigma}_U^2}$, from which we construct the series of estimators.[4] Instead of specifying a parametric form for F, we propose resampling from $\widehat{F}$, the empirical distribution for the errors.

Suppose that we were able to directly observe the errors in the variates of interest. Of course, if we actually observed all errors directly, we would not require measurement error corrections – this will be rectified through estimation shortly. Taking these errors together we can form the set $\mathcal{U}$. Sampling from $\mathcal{U}$ is then sampling from the empirical

[4] In the Laplacian SIMEX, the pseudo-errors are not drawn from the estimated error distribution, but from a complementary distribution, the form of which is derived via the parametric assumption.

distribution for the errors. As a result, sampling from this set, with replacement, allows for us to conduct nonparametric remeasurement.

It will not be sufficient to take U^* sampled from $\mathcal{U}$, and use it to replace ν[5] from the P-SIMEX procedure. However, from our previous discussions regarding the importance of the characteristic function approximating 1, we can leverage U^* in a different capacity. If we suppose that the sampling of U^* is independent of U, then the characteristic function of $U + U^*$ converges to $\varphi_U(t)^2$, as $n \to \infty$. Extending this, we can independently sample $U_1^*, U_2^*, \ldots, U_\lambda^*$ from $\mathcal{U}$, where λ is a positive integer. The characteristic function of $U + \sum_{j=1}^{\lambda} U_j^*$ converges to $\varphi_U(t)^{\lambda+1}$, as $n \to \infty$, which equals 1 when $\lambda = -1$.

If we fix Λ to be a grid of M non-negative integers, say $\{0, 1, \ldots, M-1\}$, then for any $\lambda \in \Lambda$ we can sample λ independent realizations from $\mathcal{U}$. The previous logic suggests that doing this over the grid of Λ, then extrapolating to $\lambda = -1$ according to the same procedure as the P-SIMEX will produce a valid, nonparametric measurement error correction technique. The NP-SIMEX procedure proceeds according to the following 5 steps.

1. Form the set $\mathcal{U}$.
2. Specify a fixed grid of non-negative integers, Λ.
3. For each $\lambda \in \Lambda$, $b = 1, \ldots, B$, and every i, form the variate

$$\widetilde{X}_{bi}^*(\lambda) = X_i^* + \sum_{j=1}^{\lambda} U_{bi,j}^*,$$

   where the $U_{bi,j}^*$ are sampled independently, with replacement, from $\mathcal{U}$.

4. Using $\widetilde{X}_{bi}^*(\lambda)$, compute $\widehat{\theta}_b\left(Y, X_b^*(\lambda), Z\right)$ for $b = 1, \ldots, B$. Then compute

$$\widehat{\theta}(Y, X^*(\lambda), Z) = B^{-1} \sum_{b=1}^{B} \widehat{\theta}_b\left(Y, X_b^*(\lambda), Z\right).$$

5. Fit a parametric regression model to $\{(\lambda, \widehat{\theta}(Y, X^*(\lambda), Z)) : \lambda \in \Lambda\}$ and then extrapolate to $\lambda = -1$.

[5]Recall that ν are the standard normal pseudo errors used in the remeasurement process.

4.5.1 Example Application of the NP-SIMEX

Before justifying this procedure, theoretically, we first consider extending the example in Section 4.4.2. Our goal is to estimate the fourth moment of X using $\widehat{\theta}(X) = n^{-1}\sum_{i=1}^{n} X_i^4$. To apply the P-SIMEX here, we note that, as $n \to \infty$

$$\widehat{\theta}(X^*_{bi}(\lambda_{\mathrm{P}})) \xrightarrow{p} \mu_4 + 6\mu_2(1+\lambda_{\mathrm{P}})\sigma_U^2 + E[(\widetilde{U}+\nu^*)^4].$$

Here $\nu^* = \lambda_{\mathrm{P}}^{1/2}\sigma_U\nu$, with $\nu \sim N(0,1)$, and $\mu_j = E[X^j]$. This can be further expanded to

$$\mu_4 + 6\mu_2(1+\lambda_{\mathrm{P}})\sigma_U^2 + 3\lambda_{\mathrm{P}}^2\sigma_U^4 + 6\lambda_{\mathrm{P}}\sigma_U^4 + E[\widetilde{U}^4],$$

and since $E[\widetilde{U}^4]$ is functionally independent of λ_{P}, we can take $\mathcal{G}(\lambda) = a + b\lambda + c\lambda^2$ to be the extrapolant. In this setting $a = \mu_4 + 6\mu_2\sigma_U^2 + E[\widetilde{U}^4]$, $b = 6\mu_2\sigma_U^2 + 6\sigma_U^4$, and $c = 3\sigma_U^4$. As a result, $\mathcal{G}(-1) = \mu_4 + E[\widetilde{U}^4] - 3\sigma_U^4$. This was the result stated in Section 4.4.2.

Under the assumption of normality, $E[\widetilde{U}^4] = 3\sigma_U^4$, and the P-SIMEX procedure results in consistent estimation of μ_4. If we instead considered $U_{ij} \sim t_5$,[6] then $E[\widetilde{U}^4] = 25$. Combined with $\sigma_U^2 = 25/3$, we can see that the P-SIMEX procedure leaves a residual asymptotic bias of $50/3$.

Applying the NP-SIMEX to the same problem, we wish to analyze the probability limit of $\widehat{\theta}\left(X + \widetilde{U}_\lambda\right)$, where $\widetilde{U}_\lambda \stackrel{d}{=} \sum_{j=0}^{\lambda}\widetilde{U}_j$. We find that, as $n \to \infty$,

$$\widehat{\theta}\left(X + \widetilde{U}_\lambda\right) \xrightarrow{p} \mu_4 + 6\mu_2(1+\lambda)\sigma_U^2 + (\lambda+1)E[\widetilde{U}^4] + 3(\lambda+1)\lambda\sigma_U^4,$$

which once again can be fit exactly using a quadratic extrapolant. In this case this results in $\mathcal{G}'(\lambda) = a' + b'\lambda + c'\lambda^2$, with $a' = \mu_4 + 6\mu_2\sigma_U^2 + E[\widetilde{U}^4]$, $b' = 6\mu_2\sigma_U^2 + E[\widetilde{U}^4] + 3\sigma_U^4$, and $c' = 3\sigma_U^4$. This leads to the conclusion that $\mathcal{G}'(-1) = \mu_4$, regardless of the value of $E[\widetilde{U}^4]$.

4.5.2 Theoretical Justification for the NP-SIMEX

This previous example motivates the theoretical justification for the NP-SIMEX. To justify the procedure theoretically, we demonstrate that, under a set of regularity conditions, corrections obtained through the NP-SIMEX procedure are consistent and asymptotically normal. Note that, in general, as $n \to \infty$ we know that empirical distribution functions

[6] Here we take 5 degrees of freedom to ensure that the fourth moment exists.

($\widehat{F}_X(x) = n^{-1}\sum_{i=1}^{n} I(X_i \leq x)$) converge almost surely to the true, underlying distribution function (that is, $\widehat{F}_X(x) \xrightarrow{a.s.} F(x)$). As a result, we also have that $F_{U^*} = \widehat{F}_{\widetilde{U}} \xrightarrow{a.s.} F_{\widetilde{U}}$.

Note that in the above example we treated the NP-SIMEX technique as though the estimator was computed based on random observations that are distributed as $X + \sum_{j=0}^{\lambda} \widetilde{U}_j$. In practice, we will be computing the estimator based on random quantities that are distributed as in $X + \widetilde{U}_0 + \sum_{j=1}^{\lambda} U_j^*$, where the U_j^* are sampled from $\mathcal{U}$. To justify this substitution, we take a perspective of estimands as functionals over distributions. If our interest is in θ, which is a parameter of the distribution F, then we can view $\theta = \mathbf{T}(F)$, where $\mathbf{T}$ is a functional mapping the space of distributions to the reals. Generally then it can be informative to view estimators as functionals operating on an empirical distribution, and consistency is achieved whenever $\mathbf{T}(\widehat{F}) \xrightarrow{p} \mathbf{T}(F)$.

In the case of the NP-SIMEX, we require that $\lim_{n\to\infty} \mathbf{T}(\widehat{F}) = \mathbf{T}(\lim_{n\to\infty} \widehat{F})$. As a sufficient condition, we can take $\mathbf{T}$ to be *weakly continuous*.[7] While this assumption suffices, it is not necessary; the results regarding Glivenko-Cantelli classes from van der Vaart and Wellner [92] or van der Vaart [91] can be applied instead. Under these conditions on $\mathbf{T}$, the NP-SIMEX procedure produces consistent estimators of the truth.

Theorem 4.5.1 (NP-SIMEX Consistency Theorem)**.** *Suppose that the estimator $\widehat{\theta}(X)$ can be expressed as a weakly continuous functional $\mathbf{T}(F_X)$. Moreover, assume that $\mathbf{T}(F_\lambda)$ is captured by $\mathcal{G}(\lambda)$, which has a known parametric form with parameters that are computable based on $\lambda \geq 0$ for all $\lambda \geq -1$. Under these assumptions, $\widehat{\theta}_{NP\text{-}SIMEX}$ is consistent for θ.*

In addition to consistency, under similar technical conditions on the functional $\mathbf{T}$, the limiting distribution for the NP-SIMEX estimators will be normal.

Theorem 4.5.2 (NP-SIMEX Asymptotic Normality)**.** *Suppose that the estimator $\widehat{\theta}(X)$ can be expressed as a functional, $\mathbf{T}(F_X)$, which is subject to regularity conditions such that it admits a linear approximation. Then, as $n \to \infty$,*

$$\sqrt{n}(\widehat{\theta}_{NP\text{-}SIMEX} - \theta) \xrightarrow{d} N(0, \Sigma_{NP\text{-}SIMEX}).$$

The conditions required on the functional for these two theorems are non-trivial. Weak continuity can be challenging to check, and while widely discussed, the condition that the functional admits a linear approximation is fairly technical. For instance, if $\mathbf{T}$ were Fréchet differentiable, with respect to a metric d^* such that $\sqrt{n}d^*(\widehat{F}_\lambda, F_\lambda) = o_p(1)$, where F_λ to be the distribution function of $X_i^* + \sum_{j=1}^{\lambda} \widetilde{U}_{i,j}$, then this will suffice. Alternative

[7] A functional is weakly continuous if it is continuous with respect to the weak(-star) topology [41].

characterizations are discussed in the literature. For a limited selection see Filippova [28], Kallianpur [42], Kallianpur and Rao [43], Fernholz [27], and the references therein.

Like the P-SIMEX, the NP-SIMEX requires that the functional $\mathbf{T}(F_\lambda)$ can be accurately captured by a parametric form, $\mathcal{G}(\lambda)$. This assumption is carried over from the standard SIMEX procedure, where the exact same assumption is required (though, it is typically not stated in functional language). For any specific analysis, proving the existence of such a parametric form would require in depth derivations of the underlying estimators. In practice, this assumption is less limiting than the technical regularity conditions. The rationale for this is that the modelling procedure for estimating $\mathcal{G}(\lambda)$ allows the analyst to generate arbitrary realizations in order to test the model fit. As a result, while extrapolation to $\lambda = -1$ requires faith in the existence of $\mathcal{G}$, this is the same faith required to perform any extrapolation for any model.[8]

This procedure relies on being able to form the set $\mathcal{U}$. The method for doing this depends on the auxiliary data that are available. We present methods for forming the set $\mathcal{U}$ when there is a validation sample (Section 4.5.3), and when there are replicate measurements (Section 4.5.4). When relying on replicate measurements, we require that the underlying error distributions are symmetric, but do not make specific distributional assumptions. This assumption may be reasonable since, empirically, errors often appear to follow heavy-tailed t-distributions [1, 66].

4.5.3 NP-SIMEX with a Validation Sample

Suppose that we observe an internal validation sample. That is, for some subset of individuals, say $i = 1, \ldots, n_1$ we have $\{Y_i, X_i, X_i^*, Z_i\}$, and for the remaining individuals, $i = n_1 + 1, \ldots, n$ we observe only $\{Y_i, X_i^*, Z_i\}$. In this setting, under the assumption that the validation sample is representative, then we can directly form $\mathcal{U}$ from the validation sample. Note that for any individual within the validation sample, we have $U_i = X_i^* - X_i$. If the assumed measurement error model is correct then, regardless of the distribution of U_i, this will result in the formulation of the empirical error distribution set.

If in place of an internal validation sample we have an external validation sample where we observe $\{X_i, X_i^*\}$ for $i = 1, \ldots, n_1$ and then $\{Y_i, X_i^*, Z_i\}$ for an independently sampled set of $i = 1, \ldots, n$, then we can use the external set to form $\mathcal{U}$ in exactly the same manner. Here, in addition to the assumption that the error model is correct, we must also assume

[8] One might argue that it in fact requires less faith seeing as in many situations where extrapolation is desired the analyst has a fixed quantity of data. Here new realizations can be generated limited only by the analyst's patience and computational capacity.

the standard transportability assumption that is made for studies with external validation data. If validation data are to be used, then the previously discussed convergence of the characteristic function occurs as $n_1 \to \infty$.

4.5.4 NP-SIMEX with Replicate Measurements

If instead of a validation sample we are relying on replicate measurements, we require further restrictions on the error distribution. While we do not need to make a specific distributional assumption, we do require that the distribution of U is symmetric around 0. Suppose that for all i we observe $\{Y_i, X^*_{i1}, \ldots, X^*_{ik}, Z_i\}$, each $X^*_{ij} = X_i + U_{ij}$, with U_{ij} being independent (of each other, and all other quantities) and identically distributed according to some symmetric distribution.

First, consider the case when $k = 2$. Define $\widetilde{X}^*_i = 2^{-1}(X^*_{i1} + X^*_{i2})$, which equals $X_i + 2^{-1}(U_{i1} + U_{i2})$. This can be viewed as an error-prone measurement of X_i itself. If we define $\widetilde{U}_i = 2^{-1}(X^*_{i1} - X^*_{i2}) = 2^{-1}(U_{i1} - U_{i2})$, then by symmetry we have that $\widetilde{U}_i$ will be equal in distribution to $2^{-1}(U_{i1} + U_{i2})$. Following from this, $\widetilde{X}^*_i \stackrel{d}{=} X_i + \widetilde{U}_i$. As a result, we can form $\mathcal{U} = \{\widetilde{U}_1, \ldots, \widetilde{U}_n\}$, which serves as the set to sample from for the empirical error distribution when using the mean response.

This procedure can be modified when $k \neq 2$. If we take a k-dimensional contrast, $(a_1, \ldots, a_k)$ with $\sum_{j=1}^k a_j = 0$ and $\sum_{j=1}^k |a_j| = 1$. If we consider the sums given by

$$\sum_{j=1}^k a_j X^*_j \quad \text{and} \quad \sum_{j=1}^k |a_j| X^*_j,$$

then the first sum will simplify to $\sum_{j=1}^k U_j$, while the second one becomes $X + \sum_{j=1}^k |a_j| U_j$. Owing to the symmetry of the U_j, we know that $|a_j|U_j \stackrel{d}{=} a_j U_j$, and as a result we can take $\widetilde{X}^*_i = \sum_{j=1}^k |a_j| X^*_{ij}$, and $\widetilde{U}_i = \sum_{j=1}^k a_j X^*_{ij}$, and apply the same argument as above.

In the case of $k = 2$, we have used the contrast $(1/2, -1/2)'$, which is naturally extended (when k is even) to

$$\left(\underbrace{1/k, \ldots, 1/k}_{k/2 \text{ terms}}, \underbrace{-1/k, \ldots, -1/k}_{k/2 \text{ terms}} \right)'.$$

When k is odd, we can use

$$\left(\underbrace{1/(k+1), \ldots, 1/(k+1)}_{(k+1)/2 \text{ terms}}, \underbrace{-1/(k-1), \ldots, -1/(k-1)}_{(k-1)/2 \text{ terms}} \right)' .$$

When we use replicates we replace X_i^* with $\widetilde{X}_i^*$. The assumption that U_{ij} are symmetrically distributed makes the method more restricted than in the case when validation data are available.

4.5.5 Variance Estimation

The primary drawback to the use of SIMEX procedures in general is the required computation. While the NP-SIMEX does not add computational burden compared to the P-SIMEX,[9] the process remains demanding. When considering variance estimation it is thus worth seeking alternatives to bootstrap procedures. While bootstrap procedures are valid in the context of the NP-SIMEX, the nested re-sampling adds overhead, particularly when used in a simulation, where experiments must be repeated often. There were two additional variance estimators proposed alongside the P-SIMEX, one which used a modified Jackknife procedure [87], and one which relied on the asymptotic distribution [8].

Theorem 4.5.2 allows for the use of sandwich estimation techniques to establish an estimate of the variance. The complete details are the same as any M-estimator (for instance, the complete derivation is given for the P-SIMEX by Carroll, Küchenhoff, Lombard, and Stefanski [8]). The key result is that, asymptotically, the variance of $\widehat{\theta}_{\text{NP-SIMEX}}$ can be estimated by an application of the Delta method to the estimation of the parameters of $\mathcal{G}(\lambda)$. To do this, in this context, we require an estimate for $\mathcal{C}_{11} = \text{cov}\left(\Psi_F(\Lambda)\right)$, the covariance of the stacked influence curves of the functional representation of our estimator.

While the asymptotic distribution provides a theoretically justified, large sample method for quantifying uncertainty, the primary drawback for its use in this setting is that it relies on the functional representation of the estimator. From this linearized, functional representation, a specific application of the Delta method must be used to derive the specific form for the sandwich variance estimators, which then must be programmed itself. This will generally be an involved task, mathematically, and may serve to undermine the utility of NP-SIMEX as a generally applicable measurement error correction technique.

[9] Drawing from the empirical distribution is fairly efficient, and quite similar computationally to drawing from a normal distribution.

If both bootstrap and the asymptotic distribution are not viable options, a third technique for variance estimation can be derived based on an extension of the Jackknife. To motivate this procedure, consider a sample of size 1. Suppose that the estimator of interest is expressible as a function f which is sufficiently smooth, in the sense of Stefanski and Cook [87] (which is to say, it has a convergent power series representation). Suppose that we were able to sample U_λ directly from F_λ, and consider the quantity given by $f(X^* + U_\lambda) = f(X + U_{\lambda+1})$.

The smoothness assumption allows us to write

$$f(X^* + U_\lambda) = f(X) + \sum_{n=1}^{\infty} (n!)^{-1} f^{(n)}(X) U_{\lambda+1}^n,$$

where $f^{(n)}(x)$ is the n-th derivative of f. Then, taking $E[f(X^* + U_\lambda)|X]$ we are left with

$$f(X) + \sum_{n=1}^{\infty} (n!)^{-1} f^{(n)}(X) E[U_{\lambda+1}^n].$$

Considering that, in the limit as $\lambda \to -1$, $U_{\lambda+1}$ is distributed as a degenerate distribution at 0, then all moments of the distribution are also 0. When we know the true extrapolant, have a sufficiently good estimate of the empirical distribution, and are dealing with a smooth function f, then our corrected estimator can be viewed as a conditionally unbiased estimate of $\widehat{\theta}_{\text{Truth}} = f(X)$. The notation $\widehat{\theta}_{\text{Truth}}$ refers to the estimator that would be computed if X were observable. That is, under these conditions

$$E[\widehat{\theta}_{\text{NP-SIMEX}}|X] \approx \widehat{\theta}_{\text{Truth}}.$$

If this relationship is assumed to hold exactly, then we would be able to decompose the variance of the NP-SIMEX correction as

$$\text{var}(\widehat{\theta}_{\text{NP-SIMEX}}) = \text{var}(\widehat{\theta}_{\text{Truth}}) + \text{var}(\widehat{\theta}_{\text{NP-SIMEX}} - \widehat{\theta}_{\text{Truth}}).$$

The first component of this decomposition can be estimated using an extrapolation procedure, much in the same way that SIMEX does. If, for every λ, we compute the estimated variance of $\widehat{\theta}(\lambda)$, then extrapolating this sequence of estimators back to $\lambda = -1$ presents an estimate for $\text{var}(\widehat{\theta}_{\text{Truth}})$ under exactly the same conditions that SIMEX estimates θ. For the second component of the variance, we can consider the terms $\Delta_b(\lambda) = \widehat{\theta}_b(\lambda) - \widehat{\theta}(\lambda)$, where $\widehat{\theta}(\lambda) = E[\widehat{\theta}_b(\lambda)|X]$. Note that $\text{var}(\Delta_b(\lambda)) = \text{var}(\widehat{\theta}_b(\lambda)) - \text{var}(\widehat{\theta}(\lambda))$. If we let $\lambda \to -1$, then

through a similar argument as above, this will simplify to $\text{var}(\widehat{\theta}_{\text{Truth}}) - \text{var}(\widehat{\theta}_{\text{NP-SIMEX}})$. As a result, we can use $\text{var}(\Delta_b(-1))$ as a stand-in for the second term in this variance expression. This variance term can be estimated by computing the sample covariance of $\widehat{\theta}_b(\lambda)$, which converges to its limit as $B \to \infty$. That is, for our estimator we take

$$\widehat{S}^2_{\Delta}(\lambda) = (B-1)^{-1} \sum_{b=1}^{B} (\widehat{\theta}_b(\lambda) - \overline{\widehat{\theta}}(\lambda)).$$

This argument lends itself to a SIMEX-based approach for estimating the variance. For every value of λ, we compute the average estimated variance for the underlying estimator, supposing that our remeasured variables were truth. We can then fit an extrapolant function to this sequence of estimates, and extrapolate to $\lambda = -1$. Similarly, for each value of λ, we can compute $\widehat{S}^2_{\Delta}(\lambda)$, which can also be extrapolated to $\lambda = -1$. These can be combined into a single variance estimate. In practice, we will typically form a single variance estimate for each λ given by

$$V(\lambda) = \text{var}(\widehat{\theta}_{\text{Truth}}) - \widehat{S}^2_{\Delta}(\lambda),$$

which itself can then be extrapolated to $\lambda = -1$ to approximate $\text{var}(\widehat{\theta}_{\text{NP-SIMEX}})$.

We leave the theoretical justification of this technique at this heuristic argument, which is provided in substantially more detail in Stefanski and Cook [87]. In practice, by fitting a secondary extrapolant to $V(\lambda)$, we have a computationally efficient mechanism for approximating the variance. We demonstrate the possible utility of this approach via simulation, and would advise that confirmatory simulations are used prior to the application of this technique to novel estimators. Where this technique provides unsatisfactory coverage, standard bootstrap theory or asymptotic normality can be applied, at the cost of additional computation time and additional mathematical complexity, respectively.

4.6 Simulation Studies

4.6.1 Logistic Regression Analysis

In this section we present six simulation studies, investigating the behaviour of the estimator in several scenarios. The first simulation contrasts the P-SIMEX and the NP-SIMEX in a logistic regression. We take $n = 5000$, with $B = 100$ SIMEX replicates, and with a Λ grid size of 10. We generate a true, unobserved covariate X according to a $N(1, 2)$

Table 4.3: The mean squared error (MSE) and coverage probability from 200 replicate simulations, estimating the slope parameter in a logistic regression, where the variate has t-distributed error, with varying degrees of freedom (DFs) presented. Coverage probability is computed using a bias corrected bootstrap, with 500 bootstrap resamples.

	P-SIMEX		NP-SIMEX	
DFs	MSE	Coverage	MSE	Coverage
3	0.011	0.490	0.002	0.920
4	0.014	0.215	0.002	0.915
5	0.014	0.160	0.002	0.925
10	0.015	0.125	0.001	0.935
30	0.016	0.095	0.001	0.945

distribution, and consider the outcome to be such that $P(Y = 1|X) = H(1 - X)$, where $H(\cdot)$ is taken to be the inverse-logit function. In place of X, we generate two replicated responses for each individual, X_1 and X_2, which are given by $X + U_j$, $j = 1, 2$ where U_j follows a t distribution, independent of all other variables. We take the degrees of freedom to be one of $\{3, 4, 5, 10, 30\}$. Both the P-SIMEX and NP-SIMEX are implemented using the nonlinear extrapolant, and we compute 95% confidence intervals using a bias adjusted bootstrap procedure with 500 bootstrap replicates. These simulations are repeated 200 times, due to the computational complexity of the simulations, and the results are shown in Table 4.3, where the columns under the heading MSE report the mean squared error over the 200 repeated simulations, and the columns for the coverage probability report the proportion of constructed 95% bootstrap confidence intervals which contain the true value.

We can see that, across all t distributions which were investigated, the NP-SIMEX dramatically improves over the P-SIMEX in MSE. The computed coverage probabilities are also substantially improved, though there is evidence of under coverage, particularly for low degrees of freedom. While none of these differences are significant at a 95% level, these anti-conservative results warrant caution and careful application of the bootstrap procedure, specifically when the error distribution is likely to be particularly heavy-tailed. Still, the results suggest that bootstrapping may be a feasible solution for quantifying the uncertainty in the NP-SIMEX procedure, when the computation is not a problem, so long as it has first been validated.

Table 4.4: The relative MSE from 1000 replicated simulations, estimating the fourth moment of a contaminated random variable, over different sample sizes. Values are the MSE divided by the MSE computed using the error-free covariate (truth), at the same sample size. The MSE using the true values is given.

n	Naive	P-SIMEX	NP-SIMEX	Truth
100	1.906	1.113	1.083	41160.536
500	3.975	1.264	1.087	8447.917
1000	8.275	1.759	1.339	3350.250
5000	34.404	3.561	1.451	664.606
10000	70.330	5.259	1.347	324.617
20000	133.467	8.622	1.458	170.587
50000	350.646	21.772	1.512	64.119
100000	683.371	39.205	1.490	32.579

4.6.2 Impact of Sample Size

The second simulation investigates the impact of sample size on the variability of the estimation. We use the example from Section 4.3, which involves estimating the fourth moment of X, which we take to be from a $N(5, 4)$ distribution. We take two error-prone measurements, both subjected to additive error from a t_5 distribution. The errors are independent of each other, and of the X's. We vary the sample size from 100 to 100000, replicating each 1000 times. We take $M = 10$ and $B = 500$. The MSE over the 1000 replicates when the truth is available, and the relative MSEs (that is, the observed MSE divided by the observed MSE for the true procedure) for the naive, P-SIMEX, and NP-SIMEX corrections are shown in Table 4.4.

Predictably, the naive method performs entirely unsatisfactorily, and demonstrates the utility of both the P-SIMEX and the NP-SIMEX in reducing the impacts of measurement error. While the MSE is quite large for small n, no matter the method, this is also true for the true estimator, seeing only an 8.3% and 11.3% increase in the relative MSE's over truth for the NP-SIMEX and P-SIMEX respectively (when $n = 100$). While the raw MSE decreases for both correction procedures as n increases, the relative MSE increases for both. However, the NP-SIMEX remains relatively comparable to the truth for all values of n, while for larger values of n, the P-SIMEX performs substantially worse.

Table 4.5: The MSE (multiplied by 100) from 1000 replicated simulations estimating the slope parameter in a logistic regression, over different sample sizes (n), validation sample size percentages (%), and ratios of standard deviations (σ_U/σ_X). The results compare the naive estimators, those from the P-SIMEX (P), and those from the NP-SIMEX (NP). This table contains results with a sufficiently large validation sample, relative to the measurement error variance.

	$\frac{\sigma_U}{\sigma_X} = 0.1$			$\frac{\sigma_U}{\sigma_X} = 0.5$			$\frac{\sigma_U}{\sigma_X} = 1$			$\frac{\sigma_U}{\sigma_X} = 2$		
%	N	P	NP	N	P	NP	N	P	NP	N	P	NP
						$n = 1000$						
5	1.1	1.1	1.1	22.5	11.4	12.4	–	–	–	–	–	–
10	1.1	1.1	1.2	22.0	10.8	3.9	–	–	–	–	–	–
50	1.2	1.2	1.3	22.1	10.8	2.4	72.6	26.4	8.6	–	–	–
						$n = 10000$						
5	0.2	0.2	0.1	21.9	10.3	0.9	72.6	25.9	7.1	–	–	–
10	0.2	0.2	0.1	22.0	10.3	0.7	72.6	25.7	4.8	–	–	–
50	0.2	0.2	0.1	21.9	10.2	0.6	72.4	25.4	3.5	123.0	85.6	9.2
						$n = 100000$						
5	0.1	0.1	0.0	21.9	10.2	0.4	72.5	25.4	3.3	123.0	85.8	10.0
10	0.1	0.1	0.0	21.9	10.2	0.4	72.5	25.4	3.2	123.0	85.7	7.6
50	0.1	0.1	0.0	21.9	10.2	0.4	72.5	25.4	3.2	123.0	85.7	6.8

4.6.3 Corrections with Validation Data

The third simulation considers the use of validation data in place of replicate measurements. We generate the true variate, X, to be Gamma with shape parameter 1 and scale parameter 2, such that $E[X] = 2$ and $\text{var}(X) = 4$. We generate an additive error term, U_i, which is mean-zero and follows a Laplace distribution. We consider several values for the measurement error variance, taking the ratio σ_U/σ_X to be one of 0.1, 0.5, 1, or 2. The sample size is selected to be one of $\{1000, 10000, 100000\}$, and we assume that an internal validation sample is available comprised of 5%, 10%, or 50% of the total sample. All results use the nonlinear extrapolant. The MSEs for the naive, P-SIMEX, and NP-SIMEX estimators across all scenarios are presented in Tables 4.5 and 4.6. The median squared errors for the same estimators are presented in Table 4.7.

In Table 4.5 we see that the NP-SIMEX seems to outperform both the P-SIMEX procedure and naive estimation, particularly when the ratio of variances grows. For a sufficiently small validation sample, with sufficiently small measurement error, we see that

Table 4.6: The MSE (multiplied by 100) from 1000 replicated simulations estimating the slope parameter in a logistic regression, over different sample sizes (n), validation sample size percentages (%), and ratios of standard deviations (σ_U/σ_X). The results compare the naive estimators, those from the P-SIMEX (P), and those from the NP-SIMEX (NP). This table contains results with an insufficiently large validation sample, relative to the measurement error variance.

	$\frac{\sigma_U}{\sigma_X} = 1$			$\frac{\sigma_U}{\sigma_X} = 2$		
n (%)	Naive	P	NP	Naive	P	NP
1000 (5)	728.0	294.5	991733.9	123.0	87.2	5292.5
1000 (10)	72.4	27.3	10055.5	122.8	86.0	19028.7
1000 (50)	–	–	–	123.1	86.3	19066.0
10000 (5)	–	–	–	123.0	85.9	12458.7
10000 (10)	–	–	–	123.1	85.9	5811.8

the P-SIMEX procedure performs at the same level as the NP-SIMEX. However, as the estimators stabilize, by increasing either n or the proportion of validation samples, the NP-SIMEX correction substantially outperforms either of the other methods. This table excludes results where the validation sample is particularly small, (50 or 100), with a ratio of standard deviations equal to 1, and the results where the validation sample is up to size 1000 when the ratio of standard deviations was 2. The results of these omitted scenarios are provided in Table 4.6.

These results demonstrate the instability of the nonparametric procedure at sufficiently small sample sizes, when the error is sufficiently large. Note that, as would be expected, the naive estimators are not impacted by the size of the validation sample, and the impact on the P-SIMEX is fairly small. For the P-SIMEX, the validation sample is used to estimate the variance of U_i, a process which is far more stable at small sample sizes than the nonparametric procedure used by the NP-SIMEX. The results emphasize the importance of considering the fact that, when using validation data, convergence of the correction happens in n_1 rather than n, and illustrate that if the validation sample is too small, or the estimated variation too large, nonparametric techniques may not be appropriate. Fortunately, while these results demonstrate clear instability at small sample sizes, the breakdown in performance is easy to see. We stress careful application of these techniques in settings where sample sizes may lead to instability.

To demonstrate this point consider the results summarized in Table 4.7. Here we consider the median squared error across all scenarios. We see that the relative performance

of the NP-SIMEX estimators improves dramatically across all scenarios, including those where the MSE of the NP-SIMEX estimators was discernibly worse than the other techniques. Similar improvements are seen using any truncated mean of the squared errors over the simulation results. Simply removing the largest 20 outliers (in terms of the magnitude of the MSE) for each method brings the worst performing scenario to have the MSE of the NP-SIMEX 14.8 times that of the P-SIMEX, in contrast to a ratio of 33677 with no outliers removed. These results emphasize the point that, while careful application of these techniques are required when sample sizes are small, the degraded performance of these estimators is overstated through aggregate simulation reporting since most of this decreased performance would be evident to an analyst directly investigating the results.

Table 4.7: The median squared error (multiplied by 100) from 1000 replicated simulations estimating the slope parameter in a logistic regression, over different sample sizes (n), validation sample size percentages (%), and ratios of standard deviations (σ_U/σ_X). The results compare the naive estimators, those from the P-SIMEX (P), and those from the NP-SIMEX (NP).

	$\frac{\sigma_U}{\sigma_X} = 0.1$			$\frac{\sigma_U}{\sigma_X} = 0.5$			$\frac{\sigma_U}{\sigma_X} = 1$			$\frac{\sigma_U}{\sigma_X} = 2$		
%	N	P	NP	N	P	NP	N	P	NP	N	P	NP
$n = 1000$												
5	0.5	0.5	0.4	21.9	10.6	3.0	72.5	28.4	21.2	122.9	86.6	80.5
10	0.5	0.5	0.5	21.9	10.1	1.8	72.7	26.4	12.2	122.9	86.1	64.3
50	0.6	0.6	0.5	22.0	10.3	1.2	72.5	25.9	5.8	123.0	85.8	27.9
$n = 10000$												
5	0.1	0.1	0.1	21.9	10.1	0.6	72.6	25.7	4.5	123.1	86.1	27.2
10	0.1	0.1	0.1	22.0	10.3	0.5	72.6	25.7	3.7	123.0	85.8	17.5
50	0.1	0.1	0.1	21.8	10.0	0.4	72.4	25.3	3.1	122.9	85.7	7.4
$n = 100000$												
5	0.1	0.1	<0.05	21.9	10.2	0.4	72.5	25.5	3.2	123.0	85.8	7.4
10	0.1	0.1	<0.05	21.9	10.2	0.4	72.5	25.4	3.1	123.0	85.8	6.5
50	0.1	0.1	<0.05	21.9	10.2	0.4	72.5	25.4	3.1	123.0	85.8	6.5

4.6.4 Corrections with Three Replicates

The next set of simulations extend the previous setting of estimating the fourth moment, this time assuming that there are three replicated observations. The errors are taken to

Table 4.8: The MSE and median based on three replicates, estimating the fourth moment (true value 1273) of a contaminated random variable, with either contaminated normal or normal errors. The results compare having two replicates available to the same estimators having three replicates available for the correction.

	Two Replicates		Three Replicates	
Error Distribution	MSE	Median	MSE	Median
Normal	279.076	1273.261	270.212	1272.946
Contaminated Normal ($\rho = 0.5$)	4074.825	1269.713	2011.741	1268.731

be contaminated normal distributions, with $\rho = 0$ and $\rho = 0.5$, and X remains distributed as a $N(5, 4)$ random variable. The sample size is fixed at $n = 15000$, with $B = 500$, and $M = 10$. These simulations are replicated 1000 times. These results are repeated with two available replicates. The results are shown in Table 4.8. We can see a reduction in MSE for both error distributions, with a far more substantial improvement coming when the errors are drawn with $\rho = 0.5$. In this specific context the addition of a third replicate decreased the MSE by more than increasing the sample size from $n = 15000$ to $n = 30000$ does (MSE at $n = 30000$ of 2131.141). These results lend credibility to both the proposed method for including larger numbers of replicates and demonstrate that the additional information is useful for improving the quality of the correction.

4.6.5 Jackknife Variance Estimation

In the next experiments, we investigate the proposed variance estimation technique. Taking the same scenario as in simulation 2, with n to be one of 500, 5000, 15000, or 50000, we consider using the Jackknife inspired variance estimation technique (Section 4.5.5), specifying a quadratic extrapolant for the variance terms. This extrapolant was chosen based on a visual inspection of the plots, rather than through derived theory. The simulations are replicated 1000 times, and the results are summarized in Table 4.9.

From these results we can see that this procedure tends to approximate the nominal coverage adequately, supposing that the sample size is sufficiently large. When $n = 500$ we see fairly poor coverage results, which tends to improve as n increases. It is worth reiterating that these results assumed a quadratic extrapolant for both the variance estimation and the point estimate. While this quadratic term is theoretically justified for the point estimate, the same justification was not used for the variance terms. It has been discussed that the quadratic extrapolant tended to be conservative in the standard setting

Table 4.9: The actual coverage levels, from 1000 replicated simulations, over various nominal coverage levels, with varying samples sizes using the Jackknife variance estimation technique. The estimand is the fourth moment of a contaminated random variable.

Nominal Coverage	$n = 500$	$n = 5000$	$n = 15000$	$n = 50000$
0.900	0.857	0.898	0.893	0.903
0.950	0.909	0.954	0.935	0.946
0.990	0.969	0.995	0.990	0.990

Table 4.10: The MSE of the estimates of the logistic regression parameters, across 1000 replicated simulations, comparing a naive fit, the NP-SIMEX, and the standard P-SIMEX procedure. The fit is based on a validation sample of size 5000, with asymmetric errors.

	Naive	P-SIMEX	NP-SIMEX
β_0	0.763	0.287	0.046
β_1	0.661	0.161	0.009
β_2	0.042	0.011	0.001

[16]. While this is generally advisable for a point estimate if in doubt, it is of course less desirable when estimating the variance of an estimator. Higher order extrapolants with less of a tendency to conservatively fit the data may be preferable for this purpose.

4.6.6 Non-Symmetric Error Distributions

The final set of simulations considers the use of validation data when the errors are non-symmetric. We consider the true variate to be distributed according to a Gamma distribution with shape parameter 2 and scale parameter 1. The assumed errors have shape parameter 1 and scale parameter 1.5. This gives the measurement error a slightly higher variance than the variate itself, a standard deviation ratio of $\frac{3}{2\sqrt{2}}$. The sample size is taken to be $n = 100000$, with a 5% validation sample. The true model for Y_i is a simple logistic regression, with logit link, with intercept $\beta_0 = 2.5$, slope for X_i as $\beta_1 = -1.25$, and the inclusion of an independent, standard normal variate Z_i with a slope of $\beta_2 = 1$. We repeat the simulation 1000 times, using $B = 200$, $M = 10$, and consider the nonlinear extrapolant for all parameters. The results are contained in Table 4.10.

Just as before, we see a dramatic improvement over the naive analysis when using either the P-SIMEX or the NP-SIMEX. The NP-SIMEX further improves over the P-SIMEX substantially. Note that in this case the sample size of both the full population

and of the validation sample are quite large (100,000 and 5000 respectively) which further emphasizes that the nonparametric techniques perform quite well, supposing that there is sufficient data to inform the estimation.

4.7 Further Relaxations to the Underlying Assumptions

In Chapter 3, we argued that methods which rely on replicate measurements should, wherever possible, be made resilient to the assumptions that the repeated measurements are truly, identically distributed. When the NP-SIMEX leverages validation data, these concerns are irrelevant. However, it is worth discussing the impact of non-identically distributed repeated measurements for use in the procedure. While our derivations assumed that the measurements were identically distributed for ease of exposition, this assumption was never actually used to prove the validity of the technique. Suppose that U_{ij} are symmetric, and independent, but may come from different distributions (either different families, or with different parameter values). It is still the case that, owing to the symmetry, for each j, $a_j U_{ij} \stackrel{d}{=} |a_j| U_{ij}$. As a result, when we form

$$\sum_{j=1}^{k} a_j X_{ij}^* = \sum_{j=1}^{k} a_j U_{ij} \stackrel{d}{=} \sum_{j=1}^{k} |a_j| U_{ij},$$

to form the set $\mathcal{U}$, this will be the valid empirical error distribution for

$$\sum_{j=1}^{k} |a_j| X_{ij}^* = X_i + \sum_{j=1}^{k} |a_j| U_{ij}.$$

Despite this resilience to the identically distributed assumption, the NP-SIMEX, just as with standard regression calibration and the P-SIMEX, does rely on complete replication.

If we happened to know that the U_{ij} are identically distributed, then it is possible to form a set $\mathcal{U}$ which is larger by considering permutations of the contrast a_j. Under the assumption of identically distributed errors every permutation of a_j will provide error distributions that are valid members of $\mathcal{U}$. When the errors are not assumed to be identically distributed, this same argument can be extended to the case of the given contrasts when k is even, owing to the constant multiplicative term. In the event of an odd number of replicates, where the error distributions are not identically distributed, the ordering of the contrast a_j will generally be important for determining which empirical error distribution you are working with.

Alongside considerations of whether the errors are identically distributed, we can consider the importance of symmetric errors for the case of replicate data. The utility of assuming that U_{ij} is symmetric (about zero) is that, for any a_j, we have $a_j U_{ij} \stackrel{d}{=} |a_j| U_{ij}$. Now, it is of course true that if $a_j \geq 0$, then $a_j U_{ij} \stackrel{d}{=} |a_j| U_{ij}$ trivially. As a result, supposing that U_{ij} does not follow a symmetric distribution, then so long as $a_j \geq 0$ in the given contrast, the argument still holds as written. This allows for a slight relaxation to the assumption of symmetric error distributions. Namely, so long as at least one of the repeated measurements has an error distribution which is symmetric, the NP-SIMEX can proceed by defining a contrast which has positive values for each non-symmetric entry, and still abides by $\sum_{j=1}^{k} a_j = 0$ and $\sum_{j=1}^{k} |a_j| = 1$. For instance, if there are $k-1$ repeated measurements which are subject to non-symmetric errors, with a single repeated measurement that has a symmetric error distribution, then we can define

$$\mathbf{a} = \begin{pmatrix} \frac{1}{2(k-1)} & \frac{1}{2(k-1)} & \cdots & \frac{1}{2(k-1)} & \frac{-1}{2} \end{pmatrix},$$

where the entry corresponding to the symmetric distribution is the $-\frac{1}{2}$ component.

One final consideration is that we have assumed that $U_{ij} \perp X_i$, for all i, j. While this is a common assumption, it is often the case that errors may depend on the true, underlying value. If this is the case then the presented argument for the NP-SIMEX is no longer valid. The issue is that we use the fact that, when observations are independent, the characteristic function of the sum of random quantities, is the product of their characteristic functions.

If U_{ij} are dependent on X_i, it is no longer sensible to speak of *the* empirical error distribution, as the errors associated with individual i will be drawn from a different distribution than those from individual $i' \neq i$. Conceptually, this problem can be rectified. In place of using the empirical distribution, we can instead use kernel density estimation (KDE). There have been many proposed techniques for estimating a conditional density, based on kernel methods [35]. With an estimated conditional KDE, it is possible to sample directly from this conditional distribution (see for instance Section 14.7 of Shalizi [79]).

Consider these ideas as they relate to the conditional density of $U_i|X_i$. We can view the sample, within the validation set, as observing $\{Y_i, X_i, U_i, Z_i\}$ for each individual (or

$\{X_i, U_i\}$ in the event of an external validation set). Then, we can take

$$\widehat{f}_{U|X}(u|x) = \frac{\widehat{f}_{U,X}(u,x)}{\widehat{f}_X(x)};$$
$$\widehat{f}_{U,X}(u,x) = \frac{1}{n_1 h_U h_X} \sum_{i=1}^{n_1} K_U\left(\frac{u-u_i}{h_U}\right) K_X\left(\frac{x-x_i}{h_X}\right);$$
$$\widehat{f}_X(x) = \frac{1}{n_1 h_X} \sum_{i=1}^{n_1} K_X\left(\frac{x-x_i}{h_X}\right),$$

where h_U and h_X are bandwidth parameters (selected based on the observed data), and $K_U(\cdot)$ and $K_X(\cdot)$ are kernel functions (for instance, the Gaussian kernel). Estimation of the bandwidth parameters was addressed by Hall, Racine, and Li [35], where they use cross validation based on the integrated squared error. Once estimated, the bandwidth parameters can be used to sample from the conditional distribution, given a particular value of $X = x$. Specifically, to sample conditional on $X = x$, we select an individual $i = 1, \ldots, n_1$ from the validation set weighted proportional to $K_X\left([x - x_i]/\widehat{h}_X\right)$. We then draw a realization from the K_U distribution, based on the kernel parameter $\widehat{h}_U$, centred at u_i. When using Gaussian kernels, this will result in drawing a random realization from a normal distribution with mean u_i and variance $\widehat{h}_U^2$ [79]. The analysis conducted previously can then proceed, conditional on $X_i = x_i$. The convergence of this modified procedure will be in n_1 rather than n.[10] If necessary we can also consider conditioning on additional factors (say Z_i) if those are strongly informative. While this procedure conceptually works, the difficulty is that we cannot directly condition on X_i outside of the validation sample.

Instead, we need some method for drawing from the correct error distribution, given only X_i^*. A possible technique is to repeat this procedure, using the validation sample to estimate $\widehat{f}_{X|X^*}(x|x^*)$, and then draw $\widetilde{X}_i$ based on this KDE, for each individual in the sample. This could then be used as the value of X_i to condition on. There are at least three challenges with this approach. First, there is substantial computational overhead with KDE techniques, and the added step further adds to this burden. Second, the propagation of noise throughout the estimation procedure may be a concern, particularly in small samples. Finally, it is worth questioning why we would use this procedure if we have access to $f_{X|X^*}$, in place of using this kernel technique to impute X directly.

The first issue can be overcome with suitable computational power, and is likely to be a more substantial issue for simulations and method validation, rather than for actual

[10] Which was also the case when independence was assumed.

application. The final concern is closely related to the problem of *deconvolution*, where estimates of the density of f_X are obtained through nonparametric techniques (see Chapter 12 of Carroll, Ruppert, Stefanski, and Crainiceanu [7]). Intuitively, a better estimate of the conditional density of $f_{X|X^*}$ would be required to directly impute X, rather than using it to overcome dependence in the error distribution, as is required for this procedure. In the event of an insufficient sample size, as we have seen, the estimators are unlikely to perform well regardless, which is a particular shortcoming to this nonparametric technique.

An alternative approach, which is less theoretically grounded, but more computationally feasible, is to instead draw error realizations directly from the distribution of $\widehat{f}_{U|X^*}(u|x^*)$. This procedure can proceed exactly as outlined above, and can be directly applied over the complete sample. Despite the easier application, this procedure is only going to approximately correct for dependence in the errors, even in the limit, as generally conditioning on X^* induces dependence between X and U, even where none previously existed.

To demonstrate the viability of this strategy, we consider one further simulation experiment. We consider a simulation with $X \sim N(1, 4)$, and $U|X \sim N(\rho(X-1), 1)$, where ρ is a parameter selected from $\{0, 0.5, 1, 2\}$. We take the outcome to be binomial, with $P(Y = 1|X) = \text{expit}(1 - X)$. The sample size is taken to be $n = 1000$ with a 20% validation sample, and the simulations are repeated 500 times. Within this context, we compare four different estimation strategies:

1. A standard application of SIMEX, which we expect to work well when $\rho = 0$.
2. A version of the NP-SIMEX where we first sample X from $X|X^*$, and then from $U|X$.
3. A version of the NP-SIMEX where we sample directly from $U|X^*$.
4. Sampling directly $X|X^*$, and averaging over many iterations of this.

The results of the MSE for the slope parameter estimate are contained in Table 4.11.

The results of these simulations reinforce the aforementioned discussion. For the given analysis when there is independence between U and X, the standard SIMEX estimators perform well, however, as this dependence strengthens, the corrections are less able to address concerns due to measurement error. The NP-SIMEX which fits directly to $U|X$ sees relatively comparable performance across all of the scenarios tested, and is always among the best techniques. Drawing from $U|X^*$ performs comparatively poorly when there is independence, but with dependent errors it sees a marked improvement, performing

Table 4.11: MSE of logistic regression slope parameter estimates from 500 simulation runs where there is simulated dependence between the true variate (X) and the error term (U) and the strength of this relationship is mediated by ρ.

ρ	SIMEX	NP-SIMEX ($U\|X$)	NP-SIMEX ($U\|X^*$)	$X\|X^*$
0	0.012	0.014	0.053	0.183
0.5	0.067	0.010	0.012	0.070
1	0.115	0.013	0.009	0.034
2	0.222	0.018	0.009	0.011

better than any other technique. The averaging of samples directly from $X|X^*$ tends to perform fairly well again with a strong enough dependence, though it seems unlikely to be preferable to the NP-SIMEX using $U|X^*$ or $U|X$.

The discussion of using KDEs in the event that $U_i \not\perp X_i$ may also prompt consideration of using KDEs under the assumption that $U_i \perp X_i$. Instead of forming $\mathcal{U}$ directly, we can estimate $\widehat{f}_U(u)$, and then sample from the this KDE. To do so, with equal probability we sample an index $i \in \{1, \ldots, n_1\}$ and then draw from the distribution corresponding to K_U with bandwidth parameter $\widehat{h}_U$, centred at u_i. This procedure is the *smoothed bootstrap*. There has been much written regarding the smoothed bootstrap, though it depends on the situation as to whether or not there is anything to be gained through it [20]. There is some evidence, in certain settings, that smoothing can improve the performance of estimators particularly with small sample sizes [25]. This smoothing could be applied, under the independence assumption, with either validation or repeated measurements. When the NP-SIMEX performs better through the use of this smoothing, and validation data are available, it is also possible to use the conditional KDE outlined above, even if $X_i \perp U_i$. Hall, Racine, and Li [35] demonstrate that, through their cross validation procedure, the estimated bandwidth parameter for the conditioning variables will converge to ∞ if and only if the variables are truly independent. Given a sufficiently large validation sample, sampling from the conditional KDE is equivalent to sampling from the marginal KDE.

4.8 Data Analysis

We consider an analysis of the Korean Longitudinal Study of Aging (KLoSA), following that conducted by Xu, Kim, and Li [99]. The KLoSA considers South Korean citizens, aged 45 and over, in a longitudinal study looking to determine health effects of aging. Our analysis considers data on $n = 9842$ individuals, with an internal validation sample of $n_1 = 505$, and

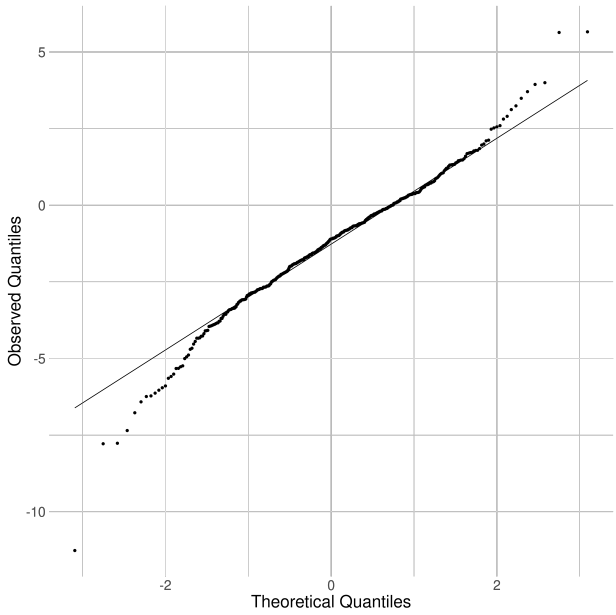

Figure 4.1: A normal Q-Q plot for the observed errors in the Korean Longitudinal Study of Aging.

we are interested in estimating how an individual's BMI impacts their propensity towards being hypertensive. In the main study BMI is estimated through self-reported weight and heights, and the validation sample includes true measurements alongside the self-reported values. Alongside the self-reported BMI, we are also given each individual's age, which we consider to be error-free.

An analysis of the validation sample demonstrates that the errors are non-normal, as is evidenced by the Q-Q plot in Figure 4.1, an excess kurtosis of 2.05, and negative skew. This suggests that the standard P-SIMEX procedure may not be appropriate. We estimate the ratio of $\frac{\sigma_U}{\sigma_X}$, using the 505 validation sample observations, as 0.898.

We analyze these data fitting a simple logistic regression model, with a logit link function, including the main effects of BMI and age. That is, we assume that

$$\text{logit}(E[Y_i|X_i, Z_i]) = \beta_0 + \beta_1 \text{BMI}_i + \beta_2 \text{Age}_i.$$

We generate bootstrap standard error estimates with 1000 replicates, and compare both the NP-SIMEX and P-SIMEX. The nonlinear extrapolant was selected for both procedures. We also consider an uncorrected analysis. The results are summarized in Table 4.12.

Table 4.12: Point estimates and bootstrap standard error (SE) estimates for the logistic regression parameters, estimating the propensity of hypertension with an intercept (β_0), self-reported BMI (β_1), and age (β_2). The estimates are based on 1000 bootstrap replicates, comparing the naive method, P-SIMEX correction, and NP-SIMEX correction.

	β_0		β_1		β_2	
Method	Estimate	SE	Estimate	SE	Estimate	SE
Naive	-5.023	0.439	0.030	0.016	0.053	0.002
P-SIMEX	-5.512	0.833	0.049	0.031	0.054	0.003
NP-SIMEX	-5.061	0.673	0.039	0.026	0.054	0.002

The three methods tend to agree on the estimate and standard error for β_2. For both β_0 and β_1, we see that the NP-SIMEX method estimates values which are larger in magnitude than the naive estimator but smaller than the P-SIMEX correction, both for the point estimate and the standard error. All three techniques suggest a positive effect of BMI on hypertension, though, the level of significance of this effect varies dramatically: 0.054, 0.117, and 0.133 for the naive, P-SIMEX, and NP-SIMEX estimators respectively.

One concern with this analysis of KLoSA is that there is strong evidence that the observed errors are not independent of the true values. This is not all together surprising, given past research findings[93]. In Figure 4.2 we can see a plot of the error terms versus the true values, illustrating the degree of dependence that is present in these data. This relationship corresponds to a correlation of approximately -0.464. With this in mind, we may question how applicable the originally provided estimators are within these data.

To supplement the previously considered analyses, we further consider conducting the same analysis using the two proposed KDE NP-SIMEX estimation techniques, based on both sampling first from $X|X^*$ and then $U|X$, and on sampling directly from $U|X^*$. The estimated coefficients and bootstrap standard errors are estimated using both of these methods, and included in Table 4.13. For each of the analyses we consider using both the quadratic and the nonlinear extrapolant. The slope coefficient for age (β_2) generally was not estimable with the nonlinear extrapolant and so these results are not reported.

The resulting estimates for β_0 and β_2 do not differ substantially from the non-conditional results. The signs for these coefficients, and their approximate magnitudes are comparable to the previously estimated values. The largest difference is in the estimates for β_1, and in particular when the nonlinear extrapolant was used.[11] These results suggest that the

[11]Note that from the theory of SIMEX we expect that generally the nonlinear extrapolant provides a better, though less conservative fit, and so the larger magnitude is not all together surprising.

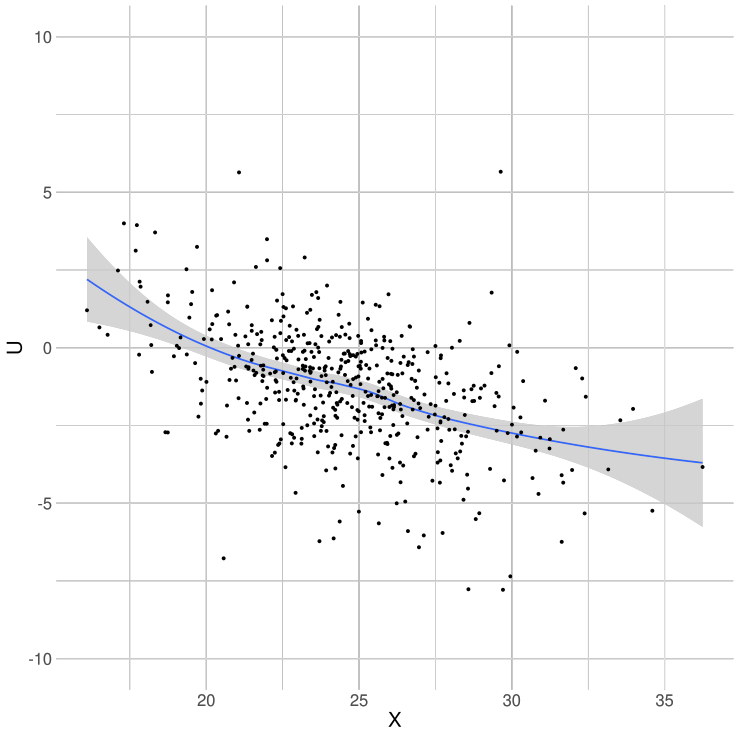

Figure 4.2: Estimated errors (U) versus the true underlying BMI for individuals within the KLoSA validation sample. The included line is a LOESS curve, included to clearly delineate the degree of dependence observed within these data.

Table 4.13: Point estimates and bootstrap standard error (SE) estimates for the logistic regression parameters, estimating the propensity of hypertension with an intercept (β_0), self-reported BMI (β_1), and age (β_2). The estimates are based on 500 bootstrap replicates, comparing the conditional NP-SIMEX method using $U|X$, and the conditional NP-SIMEX method using $U|X^*$, both with a quadratic and nonlinear extrapolant.

		β_0		β_1		β_2	
Method	Extrapolant	Estimate	SE	Estimate	SE	Estimate	SE
$U\|X$	Quadratic	-5.531	0.472	0.049	0.017	0.054	0.002
$U\|X$	Nonlinear	-4.740	0.669	0.063	0.028		
$U\|X^*$	Quadratic	-5.515	0.473	0.049	0.017	0.054	0.002
$U\|X^*$	Nonlinear	-4.936	0.679	0.061	0.038		

magnitude of the effect size was previously underestimated, quite severely. If we compare the use of the nonlinear extrapolant with either of the conditional distributions to that of the previous analyses we find that the previous estimates had magnitudes which were between 0.5 and 0.8 times the estimated magnitude using the conditional distribution. The p-values for a test of significance using $U|X$ and $U|X^*$ were respectively 0.022 and 0.105.

Given the clear dependence observed between the errors and the true BMI in these data, and in past literature, we advise taking the conditional analyses as more reliable for estimators of the truth than the unconditional analyses presented originally. Agreement on the intercept and age coefficient gives confidence in these estimates.

Part II

Dynamic Treatment Regimes

Chapter 5

Methodological Background: Dynamic Treatment Regimes

In this chapter we formally introduce dynamic treatment regimes (DTRs). We begin by discussing *potential outcomes*, which serve as a way to formalize causal inference generally. We then define a DTR in terms of the potential outcomes framework. Using this framework, we further discuss what is meant by an *optimal* dynamic treatment regime, and we introduce the necessary assumptions to identify an optimal DTR, with a causal interpretation, from observational data. We then focus on how estimation of an optimal DTR proceeds, introducing the methods of Q-learning, dynamic weighted ordinary least squares (dWOLS), and G-estimation in detail. Finally, we discuss the impacts of measurement error on the estimation of optimal dynamic treatment regimes.

5.1 Potential Outcomes for DTRs

A dynamic treatment regime is a set of decision rules that take in patient information and map to treatment decisions. We start by considering a single treatment decision. We wish to take in all information that is collected from the patient and, using this information, produce a treatment decision. This will be codified through a function $d\colon\ \mathcal{X} \to \mathcal{A}$, where $\mathcal{X}$ represents the space containing all possible patient covariate realizations X, and $\mathcal{A}$ is the set of possible treatment options. For instance, if $\mathcal{A} = \{0, 1\}$ and $\mathcal{X} = \mathbb{N}$ represents patient age, then one possible decision function is given by $d(x) = I(x < 50)$, which would assign treatment $A = 1$ if the patient were under 50 years old, and treatment $A = 0$ otherwise.

DTRs are designed to accommodate longitudinal treatment pathways. Consider a process for treating a disease where there are fixed times, $t_1, \ldots, t_K$, at which treatment decisions are required (for instance, the clinical visits in STAR*D). We allow the specific times to differ by individual, so long as each decision point j is interchangeable between individuals (in STAR*D, we do not need all follow-up visits to occur on the same day, just that "visit 1" is comparable between all subjects). We assume that, for each individual (indexed by $i = 1, \ldots, n$) we observe $(X_1, A_1, X_2, A_2, \ldots, X_K, A_K, Y)$, all without error and misclassification. Moreover, we assume that $A_j \in \{0, 1\}$ for all $j = 1, \ldots, K$.

Then, at each decision point, $1, \ldots, K$, we can define a decision function. We take $\{d_1, \ldots, d_K\}$ to map from the space of patient covariates to the set of possible treatment options. As is the case in STAR*D, the treatment options can differ at each decision point, giving spaces $\mathcal{A}_1, \ldots, \mathcal{A}_K$. Also note that, after the first decision is made, we have additional information available for the decision maker. At decision point j, we have the treatments $A_1, \ldots, A_{j-1}$, as well as the covariates $X_1, \ldots, X_j$ to inform our decision. In order to use all possible information, we take $d_j \colon \mathcal{H}_j \to \mathcal{A}_j$, where $\mathcal{H}_j$ defines the space of *patient histories* h_j. We take $h_1 = (X_1)$, and otherwise $h_j = (X_1, A_1, \ldots, A_{j-1}, X_j)$. We use overline and underline notation to refer to the past and future of a variable respectively, so that, for instance, $\overline{X}_j = (X_1, X_2, \ldots, X_j)$ and $\underline{A}_j = (A_j, A_{j+1}, \ldots, A_K)$. A K-stage DTR is a set of functions, $d = \{d_1, \ldots, d_K\}$, taking in patient histories and outputting treatment decisions at each stage. The space of all DTRs is denoted $\mathcal{D}$.

In order to develop the theory of dynamic treatment regimes, we turn to the potential outcomes framework [84, 71, 74, 68, 69]. In precision medicine, a potential outcome is a random variable that represents a patient's outcome under a pre-specified treatment. Concretely, we conceive of a random variable, $Y^{A=a}$ for every $a \in \mathcal{A}$, which represents a randomly selected patient's outcome, had they been assigned treatment a. In our binary example, $Y^{A=0}$ is the patient's outcome if they are given the control, and $Y^{A=1}$ is the patient's outcome if they are assigned the experimental treatment.

We refer to these random variables as *counterfactuals* or potential outcomes, since they represent the hypothetical outcomes that would occur if, contrary to fact, the patient received the indicated treatment (without any other changes being made). As a result, it is not generally possible to observe multiple potential outcomes for any one patient. These random variables provide a useful mechanism for discussing causal effects in the population. For instance, we may be interested in $E[Y^{A=1} - Y^{A=0}]$, called the *average treatment effect*, which gives the expected causal impact of treatment (interpreted as the change in average outcome if all individuals were given treatment $A = 1$ compared to the case when all individuals were given treatment $A = 0$). For the purpose of DTRs, we need to extend the concept of a potential outcome over multiple decision points. For a sequence of treatments

$A_1 = a_1, \ldots, A_k = a_k$, we define an individual's potential outcome, $Y^{A_1=a_1,\ldots,A_k=a_k}$, as the outcome that they would realize given that specific treatment sequence.

Now, consider a single patient selected at random from the population with history H_1, and a single-stage treatment regime, d. We define $Y^d = \sum_{a \in \mathcal{A}} Y^{A=a} I\{d(H_1) = a\}$. That is, the potential outcome under d, denoted Y^d, is the potential outcome of the treatment that would be assigned to the patient, based on their history, using d. If d has multiple treatment decisions, the potential outcomes become more complex.

To understand the necessary complexity, consider a patient that has received treatment a_1, and who is now entering the second stage of the DTR. We use X_2 to inform our choice of a_2, however, X_2 may have been impacted by a_1 itself. That is, X_2 is a random quantity that may depend on a_1. Just as with the outcome, Y, we are only able to observe one version of this random quantity, and so we consider this to be an *intermediate* potential outcome. We express this as $X_2^{A_1=a_1}$. The same will go for X_3, which will now depend on both A_1 and A_2, and so we denote $X_3^{A_1=a_1,A_2=a_2} = X_3^{\overline{A}_2=\overline{a}_2}$. This continues for the remaining stages. This results in the set of potential outcomes given by $\{X_1, X_2^{A_1=a_1}, X_3^{\overline{A}_2=\overline{a}_2}, \ldots, X_K^{\overline{A}_{K-1}=\overline{a}_{k-1}}, Y^{\overline{A}_K=\overline{a}_k}\}$.[1]

The intermediate potential outcome $X_2^{A_2}$ can be extended to accommodate a dynamic treatment regime in the same way that the single-stage Y can be. That is,

$$X_2^{d_1} = \sum_{a_1 \in \mathcal{A}_1} X_2^{A_1=a_1} I\{d_1(X_1) = a_1\}.$$

From here, between decision points $j-1$ and j, we take $X_j^{\overline{d}_{j-1}}$ to represent the potential information arising in the interval between $j-1$ and j, on account of following d to determine the treatment from decision point 1 through to $j-1$. This process can be continued through all intermediate X, until the final potential outcome Y, denoted Y^d.

The sequence of potential outcomes starts by observing X_1 as a pre-treatment covariate, and using d to inform each of the treatments from stage 1 through to the end of the treatment procedure at stage K. Following the regime produces $K-1$ intermediate outcomes, determined only by information that came before it, and a final potential outcome, Y^d, that represents the outcome of interest in the treatment procedure.[2]

[1] Technically, X_1 is not a potential outcome, as it is observed prior to any treatment. I am including it here as it is far more cumbersome to discuss "X_1 and the potential outcomes".

[2] The potential outcome framework for DTRs is notationally involved. Many authors ignore the formalization, implicitly using it instead. This will perfectly suffice for understanding this thesis. The potential outcome of a DTR is simply the outcome that would be obtained if the regime d is followed start to finish.

5.2 Optimal Dynamic Treatment Regimes

In order to estimate the optimal DTR, we must first discuss what is meant by optimality. We define the *value* of a DTR to be given by $\mathcal{V}(d) = E[Y^d]$, and we typically select Y such that higher values are preferable. Then, one regime d is considered preferable to an alternative regime $d^\dagger$ if $\mathcal{V}(d) > \mathcal{V}(d^\dagger)$. More generally, we define the optimal treatment regime d^{opt} to be given by $\arg\max_d \mathcal{V}(d)$.

Solving for the optimal treatment regime is a problem that has received considerable attention. Methods typically pose either parametric models for the potential outcomes or specify some restricted search space for d. These methods may broadly be broken down into *Q-learning*, *A-learning*, *value search*, or *classification based* approaches [90]. For this thesis, we will focus primarily on: Q-learning [98] through parametric models, a method which conducts sequential regression analyses, estimating the impact of each stage independently; dWOLS [95] which extends Q-learning using a weighted regression technique; and on G-estimation [70], a technique which relies on solving a set of sequential estimating equations. These methods are examples of Q- and A-learning, and all broadly take place within a regression framework. Prior to presenting the specific implementation for optimal DTR estimation, we discuss the necessary assumptions to draw causal conclusions.

5.3 Causal Inference and Data Assumptions

In order to interpret a dynamic treatment regime causally, we will typically require three assumptions:

1. The *Stable Unit Treatment Value Assumption* (SUTVA) [73].
2. The *No Unmeasured Confounders* (NUC) assumption [72], or the *Sequential Randomization Assumption* (SRA) [68].
3. The *positivity* assumption.

SUTVA, informally, states that if an individual receives a treatment option a, then the observed outcome for this individual, Y, is equivalent to the potential outcome $Y^{A=a}$. If $\mathcal{A}$ is a countable set, then for a patient receiving treatment a with observed outcome Y, SUTVA allows us to write

$$Y = \sum_{a^\dagger \in \mathcal{A}} I(a^\dagger = a) Y^{A=a^\dagger}.$$

SUTVA is violated if multiple versions of the same treatment exist, (for instance, a patient who knowingly received a control has a different outcome than they would have if they had unknowingly received the control), or if one patient's treatment assignment impacts another patient's potential outcomes (for instance, herd immunity in vaccine trials).

SUTVA alone suffices for endowing a causal interpretation in randomized studies. In observational studies we need to add the no unmeasured confounders or sequential randomization assumptions (for the single stage or multistage case, respectively). The NUC assumption states that all information that impacts both the treatment and the potential outcomes is measured during the observational study. That is, if X contains all recorded information prior to the treatment assignment that is measured in the study, then NUC states that $Y^A \perp A \Big| X$. In the multistage setting this assumption needs to be strengthened. The SRA can be stated as

$$\{X_2^{d_1}, \ldots, X_K^{\bar{d}_{k-1}}, Y^d\} \perp A_k \Big| H_k,$$

for $k = 1, \ldots, K$. The SRA states that, given the history up to time k (for all possible times k), all potential outcomes (past, present, and future) are independent of treatment assignment at the given time.[3]

These assumptions would be violated if any piece of information that is used to inform the treatment that is received also impacts the outcome, and is unrecorded. This assumption is not explicitly required during a randomized study since the randomization gives $Y^A \perp A$. Unfortunately, it is not possible, using the observed data, to test whether the NUC/SRA holds. Thus, causal inference based on observational data needs to be informed by subject-matter experts who are able to say whether or not it is likely to hold.

While not explicitly required to conduct causal inference generally, many methods require assuming positivity. The positivity assumption states that, for every possible treatment $a \in \mathcal{A}$, we must have that $0 < P(A = a|X) < 1$. Unlike SUTVA and NUC/SRA, positivity can be verified from the data. For instance, if both X and A are binary variables, then positivity requires that we have some observations for all of $(X = 0, A = 0)$, $(X = 1, A = 0)$, $(X = 0, A = 1)$, and $(X = 1, A = 1)$. In a sense, positivity can be viewed in light of extrapolation. We are only able to make conclusions regarding treatment sequences which could have been received, as we would require extrapolation beyond the recorded data in order to violate positivity. Making these three assumptions will suffice for this thesis. We now present the details of optimal DTR estimation.

[3]It is possible to weaken the SRA and maintain the identifiability of a DTR with causal interpretation. We will make the SRA, noting that it is stronger than is strictly necessary.

5.4 Optimal Dynamic Treatment Regime Estimation

In the following sections we present three separate estimation techniques, Q-Learning, dWOLS, and G-estimation. We begin by introducing the concept of *backwards induction*, a dynamic programming technique leveraged by each of these methods. We then present the procedures in order of increasing complexity, using the simpler methods to introduce concepts which will be leveraged in the others.

5.4.1 Backwards Induction

Backwards induction is a process for selecting the optimal decision rule by starting at the final decision point, and iteratively working backwards. Intuitively, this process of backwards induction first determines how to optimally act when all information is available, except for the final decision. Once it is clear how to make the last decision, we are able to then consider the penultimate decision. By assuming that we do act optimally after this point we have effectively encoded all relevant information into the $(K-1)$-st decision, and we can decide what an optimal treatment looks like there. We continue in this fashion until we have specified, starting with only the first piece of information, how we are to optimally act. To illustrate this, and explain why it is useful, we consider a two-stage DTR. Consider taking a randomly selected individual, with observed information (x_1, a_1, x_2, a_2), and imagine the process of determining the optimal treatment regime for them.

When they present initially, a_1 has to be decided on the basis of x_1, as $X_2 = x_2$ has not yet been observed. However, we know that a_1 is likely to influence the value of X_2, and so there is a fairly complex set of considerations to make. Imagine instead that we are at the time of the second decision, where $h_2 = (x_1, a_1, x_2)$ have all been observed. Here, in order to select the optimal A_2, we need only consider its impact on the final outcome. We choose a_2 to maximize the expected outcome, given h_2. This can be expressed in a function,

$$Q_2(h_2, a_2) = E[Y|H_2 = h_2, A_2 = a_2],$$

called the (second) *Q-function*. We can select the optimal a_2 by maximizing $Q_2(h_2, a_2)$. In particular, this gives $d_2^{\text{opt}}(h_2) = \arg\max_{a_2 \in \mathcal{A}_2} Q_2(h_2, a_2)$. We further define

$$V_2(h_2) = \max\left\{Q_2(h_2, 1), Q_2(h_2, 0)\right\}.$$

This quantity extends the previously discussed concept of our value function. $V_2(h_2)$ is interpreted as the expected value for an individual who has observed history h_2, who is

then treated optimally. Note that if $\mathcal{A}_2 \neq \{0, 1\}$ then $V_2(h_2)$ is defined as the maximum value for $Q_2(h_2, a_2)$ across all $a_2 \in \mathcal{A}_2$. The binary notation simplifies this exposition.

Taking this quantity, we can now "step backwards", and consider the first decision. In considering the choice of A_1, if the patient were assigned a treatment option $a_1^\dagger$, then we know from the preceding discussion that their expected optimal outcome becomes $V_2(x_1, a_1^\dagger, X_2)$. The only randomness in this is due to X_2 being an as-yet unobserved quantity. This suggests defining

$$Q_1(h_1, a_1) = E[V_2(x_1, a_1, X_2)|H_1 = h_1, A_1 = a_1],$$

as the first Q-function. Now, we can choose the optimal value for a_1 by the same maximization argument applied to the second Q function, which also leads to an analogous definition for $V_1(H_1) = \max\{Q_1(h_1, 1), Q_2(h_1, 0)\}$. If we define d^{opt} based on the maximization of iterative Q-functions, it can be shown that $E[Y^{d^{\text{opt}}}] = E[V_1(H_1)] \geq E[Y^d]$ for all possible treatment regimes d (see section 7.2.3 of Tsiatis, Davidian, Holloway, and Laber [90]).

If $K > 2$, this process extends as expected. We define the K-th Q-function as

$$Q_K(h_K, a_K) = E[Y|H_K = h_K, A_K = a_K],$$

the K-th optimal decision rule as $d_K^{\text{opt}} = I(Q_K(h_K, 1) > Q_K(h_K, 0))$, and the K-th value function as

$$V_K(h_K) = \max\{Q_K(h_K, 1), Q_K(h_K, 0)\}.$$

Then, starting from $j = K - 1$, and working backwards to $j = 1$, we define the corresponding Q-function as

$$Q_j(h_j, a_j) = E[V_{j+1}(h_j, a_j, X_{j+1})|H_j = h_j, A_j = a_j],$$

the optimal decision rule as $d_j^{\text{opt}} = I(Q_j(h_j, 1) > Q_j(h_j, 0))$, and the corresponding value function as

$$V_j(h_j) = \max\{Q_j(h_j, 1), Q_j(h_j, 0)\}.$$

We next introduce how this process is leveraged in Q-learning to estimate an optimal DTR. To unify notation, we will generally define $V_{K+1} = Y$. The rationale for our interest in these quantities derives primarily from the fact that the optimal DTR is characterized through the optimization of the sequential Q functions [90].

5.4.2 Q-Learning

Q- (quality) learning is a technique which was introduced in the reinforcement learning literature. Q-learning operates by specifying a model for each Q-function, and then recursively fitting these models starting at stage K and stepping backwards. While it is possible to posit any model for the Q-function, we will focus on *linear Q-learning*, where at each stage a linear regression is specified. That is, we will assume that (whenever a_j is binary),

$$Q_j(h_j, a_j; \theta_j) = \beta_j' h_j^\beta + a_j \psi_j' h_j^\psi,$$

where $\theta_j = (\beta_j, \psi_j)$ are specified regression parameters, and (h_j^β, h_j^ψ) are two (possibly identical) subsets of the history vector. The process of Q-learning is then:

1. Compute the OLS estimate of θ_K as $\widehat{\theta}_K$, by fitting $Q_K(\cdot; \theta_K)$ with Y as the outcome.
2. Specify $\widehat{d}_K^{\text{opt}} = I(\widehat{\psi}_K H_K^\psi > 0)$.
3. Define $\widetilde{V}_K(h_k) = \max\left\{Q_K(h_k, 1; \widehat{\theta}_K), Q_K(h_k, 0; \widehat{\theta}_K)\right\}$.
4. Repeat steps (1, 2, 3) for the previous stage, replacing the outcome with $\widetilde{V}$, working iteratively back to stage 1.

So long as all models are specified correctly for the Q-functions, then

$$\widehat{d}^{\text{opt}} = \left\{\widehat{d}_1^{\text{opt}}, \ldots, \widehat{d}_K^{\text{opt}}\right\},$$

will be a consistent estimator for the true optimal regime. Due to the comparative simplicity of Q-learning, it is a fairly popular technique. However, there are drawbacks that serve as motivation for the more complex methods we will discuss. In any parametric implementation of Q-learning, all models need to be correctly specified, in full, in order to have consistent estimation. This is particularly concerning for linear Q-learning. It will almost never[4] be the case that the Q-functions *can be* linear, and even less often the case that they *will be*. The issue of non-linearity can of course be overcome by considering a more flexible class of models, however, the problem of complete specification remains. Non-parametric techniques can rectify these issues, at a far greater computational burden.

[4] A straightforward derivation shows that if treatment interacts with a continuous covariate, this will tend to make the earlier stages nonlinear, due to the indicator function in computing $Q_.$.

Since Q-learning fits a sequence of least squares estimators, it may seem reasonable that the standard theory of M-estimation could be exploited to derive asymptotic standard errors and distributional results. Unfortunately, that is not the case, and will generally not be the case for any DTR estimation. The issue stems from the use of the maximization operation, which is generally not differentiable. This makes the Q-learning estimator a so-called *non-regular estimator*, in the sense that it does not conform to the regularity conditions required for standard inference [70]. The central issue is that non-regular estimators may not have a unique limiting distribution. In Appendix C we give a concrete example of why this occurs. Owing to this non-regularity, standard bootstrap procedures are also unlikely to be entirely theoretically justified. Despite this, the use of M-estimation as a guiding framework remains effective. The consistency results do not rely on the regularity of the estimators, and whenever optimal treatments are clearly defined (in the sense that the counterfactual outcomes are substantially different between the different treatment options, for all individuals), standard asymptotic theory can be leveraged.

Dynamic weighted ordinary least squares, presented next, extends Q-learning in such a way to provide robustness to model misspecification.

5.4.3 Dynamic Weighted Ordinary Least Squares

In principle, dWOLS is a similar technique to Q-learning. In dWOLS we once again fit a series of regression models, using backwards induction. However, dWOLS modifies linear Q-learning in that the regressions are weighted. This weighting results in a significantly more robust methodology. To understand why, consider the linear model that was posited for Q_j, given by $\beta_j' h_j^\beta + a_j \psi_j' h_j^\psi$. The expression has been carefully constructed to indicate that the assigned treatment is only relevant through the second term. We have divided the model for Q_j into a component which captures the effect of treatment, and one which is free of treatment effects. We refer to the latter as the *treatment-free* component.

Define $\gamma_j(h_j, a_j; a_j^{\text{ref}}) = Q_j(h_j, a_j) - Q_j(h_j, a_j^{\text{ref}})$. Then $\gamma_j(h_j, a_j)$ represents the difference in the expected (optimal) outcome for a patient with history h_j, between receiving treatment a_j and a_j^{ref}. Plugging in previous definitions, we have

$$\gamma_j(h_j, a_j) = E\left[Y^{\overline{a}_{j-1}, a_j, \underline{a}_{j+1}^{\text{opt}}} - Y^{\overline{a}_{j-1}, a_j^{\text{ref}}, \underline{a}_{j+1}} \middle| H_j = h_j \right].$$

This compares the expected outcome for two individuals who have identical history up to stage j, where one of the individuals receives treatment a_j and the other receives a_j^{ref}, before they both go on to receive optimal (though, possibly not identical) treatment. We

call this quantity the *blip-to-reference function* (or more succinctly, just the blip function).

A positive blip function means that a_j is a preferable treatment to a_j^{ref}, given the history. Since we consider $A_j \in \{0, 1\}$ by assumption, we typically take $a_j^{\text{ref}} = 0$. Here the zero level is defined as our reference category, for instance, as the control in a clinical trial. Then, we would have that $\gamma_j(h_j, 0) = 0$ for all j, and that $\gamma_j(h_j, 1)$ represents the impact of receiving treatment $A_j = 1$, if everything else remains optimal. The blips define the sequence of optimal treatment rules. If $\gamma_j(h_j, 1) > 0$, then selecting $a_j = 1$ is preferable to selecting $a_j = 0$. As a result, $d_j^{\text{opt}} = I(\gamma_j(h_j, 1) > 0)$.

We can also use the blips to re-write our models for the Q-functions, taking

$$Q_j(h_j; \theta_j) = \beta_j' h_j^\beta + \gamma_j(h_j; \psi_j).$$

In Q-learning, any model misspecification will possibly lead to inconsistent results, whether the misspecification is actually encoding the impact of treatment or not. The following theorem illustrates that this is not true of dWOLS (Wallace and Moodie [95] Theorem 1).

Theorem 5.4.1 (dWOLS Consistency Result)**.** *Assume that*

$$Q_K(h_K, a_K) = f(h_K^\beta; \beta) + a_K \psi_K' h_K^\psi,$$

for some function f (functionally independent of the treatment). Define

$$\pi(h_K) = P(A_K = 1 | H_K = h_K),$$

and define $w(a_K, h_K)$ to be a weight function such that

$$\pi(h_K) w(1, h_K) = (1 - \pi(h_K)) w(0, h_K). \tag{5.4.1}$$

Under correct specification of h_K^ψ, a weighted ordinary least squares regression of the outcome Y on $\{h_K^\beta, a_K h_K^\psi\}$, using weights given by $w(a_K, h_K)$, will consistently estimate ψ_K.

This theorem indicates that the simple process of weighting, with weights satisfying Equation (5.4.1), makes the correct specification of the treatment-free component unnecessary for consistent estimation. The identity that the weights need to specify does lend some flexibility into the precise form, though it is advised that weights of the form $w(a, h_j) = |a - E[A_j | H_j = h_j]| = |a - \pi(h_j)|$ are used [95, 52]. This works via *covariate balance*, where in the weighted data set, the covariates and treatment behave as though they are independent. Any weights satisfying the requirements will rely on the propensity of treatment, $\pi(H_K)$. In the event that these propensities are readily available (for instance,

in the case of a randomized trial), we can then be certain that our blip parameters are consistently estimated, so long as the form of the blip is correctly specified. If $\pi(\cdot)$ are not known, it may be the case that this is a quantity that can be modelled easily.

If weights satisfying Equation (5.4.1) can be consistently estimated by fitting a parametric model to $P(A_K = 1|H_K = h_K; \alpha_K)$, then using estimated weights maintains the result of Theorem 5.4.1. We refer to this additional parametric model as the *treatment model*. The specification of a treatment model makes dWOLS a *doubly-robust* estimation technique. Assuming the blips are correctly specified, if either the treatment model or the treatment-free model are correctly specified, then the parameters ψ_K will be consistently estimated. This weighted regression forms the basis of dWOLS. Until now, we have only discussed the K-th stage parameter estimates. In order to estimate the blip parameters for stages $j = 1, \ldots, K-1$, we need to introduce the concept of *regrets* [59].

We can frame the observed outcome Y as the outcome under the optimal treatment regime, $Y^{d^{\text{opt}}} = Y^{\text{opt}}$, minus the impact of all suboptimal treatments that were received. Regrets are the functions which quantify the impact of suboptimal treatment. Notationally this is expressed as $\mu_j(h_j, a_j) = E\left[Y^{\overline{a}_{j-1}, a_j^{\text{opt}}, \underline{a}_{j+1}^{\text{opt}}} - Y^{\overline{a}_{j-1}, a_j, \underline{a}_{j+1}^{\text{opt}}} \middle| H_j = h_j\right]$. Just as with the blip functions, the regrets can be used to determine the optimal treatment, as we know that $\mu_j(h_j, a_j) \geq 0$, with equality only at the optimal treatment. We can also write that $\mu_j(h_j, a_j) = \gamma_j(a_j, a_j^{\text{opt}}) - \gamma_j(h_j, a_j)$. Re-writing Y using regrets can be done as

$$E\left[Y \middle| H_K = h_K\right] = E\left[Y^{\text{opt}} \middle| H_K = h_K\right] - \sum_{j=1}^{K} \mu_j(h_j, a_j).$$

This motivates a type of *pseudo outcome* for dWOLS. The basic idea is that, at each stage, we wish to estimate the outcome under the observed treatment up to the current stage, assuming that the patient goes on to receive optimal treatments at all future stages. At stage $K-1$ this is given by $Y + \mu_K(h_K, a_K)$. Generally, for any stage $j = 1, \ldots, K-1$, we take $\widetilde{Y}_j = Y + \sum_{\ell=j+1}^{K} \mu_\ell(h_\ell, a_\ell) = \widetilde{Y}_{j+1} + \mu_{j+1}$. We can estimate these using the estimated regrets in a backwards induction process. Taken together, dWOLS is summarized through the following procedure:

1. Specify a parametric model, divided into the treatment-free (indexed by β_j) and blip (indexed by ψ_j) components, for each stage (where we are modelling the expected outcome assuming that the patient goes on to receive optimal treatment).

2. Specify a parametric model for the propensity of treatment (indexed by α_j) for each stage.

3. Using the parametric treatment model, compute the weights for each individual, given by $w(a_j, h_j; \alpha_j)$, so that they conform to Equation (5.4.1).

4. Starting at stage K and working backwards, compute the estimated pseudo outcome $\widetilde{Y}_j = Y + \sum_{\ell=j+1}^{K} \mu_\ell(h_\ell, a_\ell; \widehat{\psi}_\ell)$.

5. With the computed pseudo outcome, fit the model to generate estimates for ψ_j, and then repeat the regression procedure.

With this procedure, so long as, at every stage, either the treatment-free or the treatment model is correctly specified, along with all of the blip functions, the resultant estimators are consistent. In certain circumstances, the treatment model will be known exactly. Even when it is not, it is likely the case that these experts have a good sense of the factors that are used to inform treatment decisions. This added robustness is an attractive feature shared by G-estimation, which is presented next.

5.4.4 G-Estimation

Like both Q-learning and dWOLS, G-estimation [70] is a sequential, model fitting procedure.[5] For the process of estimating d^{opt}, G-estimation begins with a consideration of the Q and value functions introduced in Section 5.4.1. If we take $A_j \in \{0, 1\}$, we can express $Q_j(H_j, A_j) = \nu_j(H_j) + A_j C_j(H_j)$, for some arbitrary functions ν_j and C_j. Then, the value function will take on the value of $\nu_j(H_j)$, if $C_j(H_j) \leq 0$ and $\nu_j(H_j) + C_j(H_j)$ otherwise. Note that in the context of G-estimation we are using notation which differs slightly than that discussed for dWOLS. In dWOLS we used $\gamma_j(\cdot)$ to denote the term given by $C_j(\cdot)$. The reason for this difference is two-fold. First, G-estimation does not implicitly restrict $C_j(\cdot)$ to take a linear form, while dWOLS does restrict $\gamma_j(\cdot)$. Second, the notation $C_j(H_j)$ has been used in recent literature surrounding G-estimation (see Tsiatis, Davidian, Holloway, and Laber [90]) while $\gamma_j(\cdot)$ is common for dWOLS.

G-estimation belongs to a class of estimation techniques known as *A-learning* (A for advantage). These methods use the insight exploited when introducing dWOLS that only the contrast (or blip) needs to be estimated in order to define the optimal treatment. The optimal treatment at each stage is determined solely by $C_j(H_j)$. If $C_j(H_j) > 0$ then

[5] The discussion of G-estimation is taken not directly from Robins [70], but rather from Wallace and Moodie [95] and Tsiatis, Davidian, Holloway, and Laber [90], owing to the clearer formulation. The Robins's paper is a sprawling account, which provides rigorous theoretical justification, but which is not particularly informative for a first-time reader.

$A_j^{\text{opt}} = 1$, and $A_j^{\text{opt}} = 0$ otherwise. Combining this result with the fact that finding the optimal DTR is equivalent to optimizing the Q functions, A-learning methods proceed by estimating $C_j(H_j)$ from the available data and taking $d_j(H_j) = I(C_j(H_j) > 0)$ to assign treatment. The term $C_j(H_j)$ is exactly the blip function from dWOLS when the reference level is taken to be the zero. As a result, we introduce the parameter ψ_j to index these functions, so that $Q_j(H_j, A_j) = \nu_j(H_j) + A_j C_j(H_j; \psi_j)$. We will assume that each C_j is correctly known up to the blip parameter indexing it. In this setting ν_j corresponds to the treatment-free model from dWOLS.

Assuming that the *treatment assignment probabilities*, $P(A_K = 1|H_K) = \pi_K(H_K)$, are known, Robins [70] demonstrated that every consistent and asymptotically normal estimator for ψ_K will solve

$$\sum_{i=1}^{n} \lambda_K(H_{i,K}) \left\{A_{i,K} - \pi_K(H_{i,K})\right\} \left\{Y_i - A_{i,K} C_K(H_{i,K}; \psi_K) + \theta_K(H_{i,K})\right\} = 0. \tag{5.4.2}$$

Both $\lambda_K(\cdot)$ and $\theta_K(\cdot)$ are taken to be arbitrary functions of the history, where $\lambda_K(\cdot)$ is constrained to be the same dimension as ψ_K. If we define $\widetilde{V}_{K+1} = Y$, and for $j = 1, \ldots, K$,

$$\widetilde{V}_j = \widetilde{V}_{j+1} + (A_j^{\text{opt}} - A_j) C_j(H_j; \widehat{\psi}_j), \tag{5.4.3}$$

then $E[\widetilde{V}_{j+1}|H_j, A_j] = Q_j(H_j, A_j) = \nu_j(H_j) + A_j C_j(H_j)$, almost surely. This holds almost surely so long as $\widehat{\psi}_j$ is almost surely consistent for ψ_j. Based on these *pseudo outcomes*, we can extend the estimating equations in Equation (5.4.2) to all stages, $j = 1, \ldots, K$ by replacing Y_i with $\widetilde{V}_{i,j+1}$. We take,

$$U_j(\psi_j) = \sum_{i=1}^{n} \lambda_j(H_{i,j}) \left\{A_{i,j} - \pi_j(H_{i,j})\right\} \left\{\widetilde{V}_{i,j+1} - A_{i,j} C_j(H_{i,j}; \psi_j) + \theta_j(H_{i,j})\right\}. \tag{5.4.4}$$

Then, solving $U_j(\widehat{\psi}_j) = 0$ renders $\widehat{\psi}_j$ a consistent estimator for ψ_j.

Just as with dWOLS, it will often be the case that we do not exactly know the treatment assignment probabilities, $\pi_j(H_j)$. If it is possible to specify a parametric model for these treatment probabilities, indexed by α_j, which can be framed as the solution to a set of unbiased estimating equations, $U_{\text{trt},j}(\widehat{\alpha}_j) = 0$, then we can simply "stack" $(U_{\text{trt},j}(\alpha_j)', U_j(\psi_j, \alpha_j)')'$ together and jointly solve them. This will result in consistent estimators when both models are correctly specified. We will often use logistic regression for the purpose of estimating the treatment assignment probabilities when they are unknown.

The functions $\lambda_j(H_j)$ and $\theta_j(H_j)$ can both be arbitrarily selected, so long as the dimension of $\lambda_j(\cdot)$ matches that of ψ_j. Typically, $\theta_j(H_j)$ can be selected to add robustness to the estimation procedure. If we are able to specify a parametric form for the nuisance parameter $\nu_j(H_j)$, indexed by β_j, then by taking $\theta_j(H_j;\beta_j) = -\nu_j(H_j;\beta_j)$ this estimation procedure becomes doubly robust in exactly the same sense that dWOLS is doubly robust. Just as with the treatment assignment probabilities, these *treatment-free* models are unlikely to be known directly. Instead, if the parameter can be estimated as the solution to $U_{\text{tf},j}(\widehat{\beta}_j) = 0$, then we can further "stack" $(U_{\text{tf},j}(\beta_j)', U_{\text{trt},j}(\alpha_j)', U_j(\psi_j, \alpha_j, \beta_j)')'$ to jointly estimate $(\beta_j', \alpha_j', \psi_j')$. The estimators for ψ_j are consistent so long as the blip model, $C_j(H_j, \psi_j)$, is correctly specified, in addition to at least one of the treatment or treatment-free models.

While the specification of $\theta_j(H_j)$ was selected to add robustness, the selection of $\lambda_j(H_j)$ is typically made with efficiency in mind. Robins [70] provides the optimal form for $\lambda_j(H_j)$, based on the asymptotic variance of the estimators, but the form is often complicated. Instead, we typically define $\lambda_j(H_j) = \frac{\partial}{\partial \psi_j} C_k(H_j;\psi_j)$, which seems to work well in practice [14, 90]. With these functions specified, G-estimation then estimates the optimal treatment regime, d^{opt} by recursively estimating the parameters ψ_k for $k = 1, \ldots, K$, and then taking $\widehat{d}_k^{\text{opt}}(H_k) = I(C_k(H_k;\widehat{\psi}_k) > 0)$.

We once again start at decision point K, and then:

1. Specify a model for the treatment-free component, $\theta_K(H_K;\beta_K) = -\nu_K(H_K;\beta_K)$.
2. Specify a model for the treatment probabilities, $\pi_K(H_K;\alpha_K)$.
3. Specify a model for the blip/contrast, $C_K(H_K;\psi_K)$.
4. Jointly solve models (1)–(3), giving estimates for ψ_K, using Equation (5.4.2).
5. Compute the pseudo outcome, $\widetilde{V}_k$ according to Equation (5.4.3), and then repeat the previous steps, solving according to Equation (5.4.4) for stages $k = 1, \ldots, K-1$.

The optimal decision rules, just as with dWOLS, can then be inferred from the estimated blip parameters. The robustness of G-estimation, and the flexibility to accommodate non-linear modelling strategies, makes it an attractive procedure for estimating optimal DTRs. The main downside of the methodology stems from the complexity, making description and implementation more cumbersome than the previously specified techniques.

5.5 Measurement Error in DTRs

Prior to this dissertation, to the best of our knowledge, there has been only one substantive attempt to assess, and correct for, the issues of measurement error in the context of DTRs. This work was completed as my masters project. This established the impact of measurement error on the estimation of optimal DTRs, illustrating the complexity of errors in these models. Assume that, for a scalar-valued, continuous, tailoring covariate, we observe classical additive error, with normally distributed errors. To assess what impact this error has on the estimation of an optimal DTR, we will use the dWOLS procedure. The impact of errors in regressor variables has been thoroughly studied, and given the strong ties of dWOLS to ordinary regression, we will expect similar issues to arise. The impact of measurement error in a DTR is complicated by the fact that the treatment, treatment-free, and blip models are all impacted by mismeasurement separately.

We assume that a biological process (or similar) relates the true covariate values X to the outcome Y, whereas treatment decisions can only be made based on the observed values X^*. This structure is shown in Figure 5.1. Because of the complicated structure, we consider the impact in the treatment, treatment-free, and blip models separately. When using dWOLS, correcting for error in covariates within the outcome model is the same as correcting for the effects of error in a standard linear regression. In this setting, we can use (for instance) regression calibration to consistently estimate all necessary parameters.

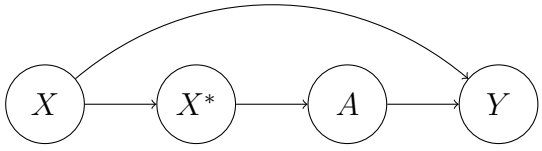

Figure 5.1: A directed acyclic graph (DAG) representing the assumed impact of measurement error in the DTR setting. Here X is the error-free covariate, X^* the observed proxy, A is a binary treatment indicator, and Y is a numeric outcome of interest.

The treatment model is leveraged primarily as a means of endowing the dWOLS estimators with the property of double robustness. It is possible to think of this double robustness as arising through the property of covariate balance [95]. In the error-prone setting, if we employ regression calibration using $\widehat{X}$ in our outcome models, we wish to induce covariate balance not between X and A, but between $\widehat{X}$ and A. Following the justification of the choice of weights by Wallace and Moodie [95], we might speculate that any weights which satisfy $\pi(\widehat{X})w(1,\widehat{X}) = (1-\pi(\widehat{X}))w(0,\widehat{X})$ will induce covariate balance

in $\widehat{X}$. During my masters project, I explored this line of reasoning and demonstrated that a small-sample proxy for covariate balance, *sample balance*, arises through these weights.

Result 5.5.1. *If* $P(A = 1|X = x) = H(\alpha_0 + \alpha_1 x)$*, where* $H(x) = (1 + \exp(-x))^{-1}$ *is the inverse-logit (expit) function, and* $X_j^* = X + U_j$*, for* $j = 1, \ldots, k$ *are iid replicate measurements of* X*, then denoting the regression calibration estimates* $\widehat{X}$ *and using the weights* $w(A, \widehat{X}) = |A - \pi(\widehat{X})|$*, where* $\pi(\widehat{X})$ *is the estimated probability* $P(A = 1|\widehat{X})$ *fit using logistic regression guarantees* sample covariate balance. *That is,*

$$\frac{\sum_{i=1}^n w(1, \widehat{X}_i) A_i \widehat{X}_i}{\sum_{i=1}^n w(1, \widehat{X}) A_i} = \frac{\sum_{i=1}^n w(0, \widehat{X}_i)(1 - A_i)\widehat{X}_i}{\sum_{i=1}^n w(0, \widehat{X})(1 - A_i)}.$$

The quality of using sample balance as a small scale proxy was assessed using simulations. This result provides a reasonable justification for using the regression calibration correction in the treatment model.

The final component that needs to be considered to address the impacts of errors in variables relates to the construction of pseudo outcomes. During the previous work, the problem of constructing a valid pseudo outcome was put aside, relying instead on justification via simulation. As a result, the existing literature on addressing measurement error in DTRs primarily served as a mechanism for demonstrating that (as one would expect) errors cause concerns for correctly estimating the optimal DTR and that these concerns are generally complex owing to the structure of a DTR. Some simulation evidence demonstrated that regression calibration may be an effective tactic for restoring desirable properties of the dWOLS estimators, but these results were explored primarily based on heuristics rather than being thoroughly justified theoretically. Moreover, the project focused on a fairly limited scope (considering only scalar-valued variables, with normal additive error). No attention was given to uncertainty quantification, either through confidence intervals or standard error estimates.

Starting from a position wherein these problems have been identified, in Chapter 6 we expand on these ideas, giving theoretical justification for the application of regression calibration to the correction of the effects of errors in DTRs. We demonstrate how, using ideas from Chapter 3, the expanded regression calibration correction can be applied in this setting. We propose the use of a modified bootstrap procedure to produce confidence intervals. These results are justified theoretically, through simulation experiments, and with an application to the analysis of STAR*D.

5.6 Nonadherence in DTRs

While the effects of errors in tailoring variables on the estimation of optimal DTRs has been briefly explored prior to this thesis, the impact of misclassification of treatment indicators, known better as *nonadherence*, has not been. There is some literature which discusses nonadherence in the context of dynamic treatment regimes. However, this research has focused on the ability to adequately estimate the expected impact of a regime that might not have been complied with, when that particular regime is of primary interest, and compliance information is readily available [39, 18, 36]. The concern in this literature is not predominantly estimating the optimal regime taking into account nonadherence, but instead, on how to identify the *value* of a particular regime of interest, when it may not have been adhered to perfectly. Wallace [94] briefly explores nonadherence in binary treatment indicators for simple dynamic treatment regimes. In this work it is shown that nonadherence can have a sizable impact on the estimation of optimal dynamic treatment regimes, even in simple examples. While the bias that nonadherence introduces is indicated and justified, no corrections are proposed.

Beyond the specific context of DTRs, nonadherence occurs when a prescribed treatment is not followed by an individual. This is a concern in the medical setting as health outcomes can be severely impacted by nonadherence. Medical researchers have been concerned with assessing the degree of nonadherence, the impacts of nonadherence on patient outcomes, the factors that predict nonadherence, and ways to limit nonadherence [103, 22, 33, 37]. From a statistical perspective nonadherence is a concern as, when present in data, it can bias parameter estimates or change the causal interpretation of parameters.

Any analysis which proceeds by ignoring nonadherence is referred to as an *intention-to-treat* (ITT) analysis [54, 34]. When conducting an ITT the causal effect that is being estimated is treatment prescription, rather than treatment itself. An ITT analysis can be motivated by taking a policy perspective. If a policy is being considered which would prescribe some recommended treatment for every patient with a particular illness, then the causal effect of this policy can be decomposed into two components. First, such a prescription likely has an impact on patient behaviour and their adherence to the prescribed treatment. Second, the treatment that the individual ends up taking has some causal effect on their outcome through relevant underlying processes.

In a clinical setting, intention-to-treat analyses are often framed in contrast to *per-protocol* (PP) analyses [67]. In a PP analysis, only the patients who adhered to the assigned treatment are included. A concern with a PP analysis is that bias will be introduced into parameter estimates if nonadherence is not completely at random. For instance, if, as

is likely the case, the patients who fail to adhere to their prescribed treatments are the ones who were not improving under the treatment, then excluding these cases from an analysis will overestimate the positive impact of treatment. An ITT analysis circumvents this source of bias by taking the act of prescription to be the intervention, rather than the treatment itself. While the concerns with PP analyses are certainly valid, there are several issues with ITT analyses as well. Of primary concern, an ITT does not provide an unbiased estimate of the effectiveness of treatment itself [81].

In Chapter 7 we provide the first procedure for estimating an optimal DTR using data which is subject to nonadherence. In addition to providing a modified version of G-estimation to account for the effects of nonadherence, we deeply explore the specific impacts of nonadherence in this context, and draw attention to the complexities that arise from this setting. We pay particular attention to the fact that, especially in the context of precision medicine, the drawbacks of an ITT should not be overlooked. We explore the ways in which nonadherence can undermine the causal interpretations that are typically made when estimating an optimal DTR, and argue that in many settings the proposed treatment-efficacy approach will be more appropriate. These results are presented theoretically, explored via simulation studies, and demonstrated through an analysis of the Multicenter AIDS Cohort Study (MACS).

5.7 Measurement Error and Nonadherence in STAR*D and MACS

While we contend that measurement error and nonadherence are likely far-reaching problems in the realm of DTR estimation, we content ourselves with an exploration of these ideas using STAR*D and MACS in this thesis.

STAR*D provides an interesting use case for the methods explored in Chapter 6, as it has never been analyzed accounting for errors in measurement, previously. We are primarily concerned with each patient's QIDS score, and as noted in Chapter 1, this was assessed at each visit both by the patient themselves and by the clinician. The self-reported QIDS score tends to be ignored in analyses of the STAR*D data, under the implicit assumption that QIDS-C is an error-free measurement of the true depressive symptomatology. If we are willing to view both QIDS-C and QIDS-S as measurements of the truth, then the presence of two separate measurements constitutes auxiliary data which can be used to correct for the effects of measurement error. We will again assume a non-differential error mechanism, modelled classically, using structural methods.

MACS, on the other hand, exists in a domain which has more explicitly handled non-adherence previously. Despite this, when analyzed as a DTR, MACS has never had this information taken into account. Generally our analysis will be concerned with the timing of AZT prescription, as it is mediated by several patient characteristics. It is well-known that patients may be nonadherent to their AZT prescription, and as a result, in some waves of the study, the researchers running MACS have collected adherence information. It is our interest in considering what the impact of this estimated degree of nonadherence will be on the assessment of treatment efficacy.

Chapter 6

DTRs with Errors in Tailoring Covariates

6.1 Motivation for Error Corrections in DTRs

The previous work conducted on errors in the tailoring variates of DTRs makes clear that a "naive" analysis, one which ignores errors when they are present, predictably leads to biased and inconsistent estimators for the true population parameters. This fact alone is not necessarily sufficient to motivate the development of measurement error correction techniques in this setting. The importance of developing methods to address the concerns of measurement error depends largely on the goals of the underlying analysis. If it is of primary interest to quantify the effect of treatment, or a treatment-covariate interaction, then the existing results demonstrate that error correction techniques are necessary.

However, it may be the case that a DTR is estimated explicitly for the purpose of future treatment assignment. In principle, the problem of assigning future treatments can be framed as a prediction problem. In this case, the DTR models are fit in service of generating optimal predictions, based on observed covariates, out of sample. Then, if it is likely that future individuals will also have their relevant tailoring factors measured with error, it may seem reasonable to ignore the errors all together. In this context we are re-specifying the problem of interest to be "predict the optimal treatment based on the surrogate measurement of the variable of interest".

The use of surrogate measurements in prediction problems, ignoring the impacts of measurement error, is a topic which has been explored in some detail. This strategy

will sometimes lead to acceptable performance, but this is not always the case. Even in comparatively simple models it is often the case that the effects of measurement error are worth correcting when the primary goal is prediction. Schaalje and Butts [77] demonstrate in the case of linear regression, where prediction is of primary interest, there are settings where ignoring the error leads to undesirable results. Khudyakov, Gorfine, Zucker, and Spiegelman [46] draw similar conclusions for risk prediction models, in a medical context, where the models are fit using GLMs. Given the complexity of predictions within the DTR setting, these issues are worth investigating further.

When approaching corrections for the impacts of measurement error, it is instructive to consider how the different aspects of a dynamic treatment regime may be impacted. As was introduced in Section 5.5, previous work on the effects of measurement error on optimal DTR estimation has been framed through the different model components. The complex and unique structure of this error (Figure 5.1) lends itself to this type of piece-wise analysis. For a dynamic treatment regime estimated through dWOLS, tailoring variables play three distinct roles. First, the tailoring variables act as predictors, used for both the treatment-free and blip components of the outcome model. Second, the tailoring variables may act as predictors in the treatment model, which is commonly a logistic (or probit) regression. Finally, the tailoring variables are used to construct the pseudo outcome, allowing backwards induction to proceed.

With the outcome model the goal is to estimate the true, underlying relationship between the tailoring variable, free from error, and the outcome. As a result, any method which corrects for the effects of measurement error in parameter estimation within linear models can potentially be used to remedy the impacts of measurement error on the first component. This would include all of the strategies previously discussed, among others. Since dWOLS leverages linear outcome models, there are an abundance of correction techniques that are suited to restoring the consistency of the estimators. We will focus on regression calibration, though conceptually, simulation extrapolation would work as well.

There exist many methods for correcting for errors in predictors in a logisitc regression model (see for instance the discussion of moment reconstruction in Chapter 3). The goal of these corrections is to model the true relationship between the underlying, error-free variable and the outcome. When we consider the treatment model, the goal is not to model the true relationship between the tailoring covariate and the treatment indicator, but instead to produce weights which endow the resultant estimators with double robustness. As a result, when selecting a technique for addressing the errors in the treatment models, our considerations are not around consistent estimation of the true parameters, but rather, around the weights meeting the necessary conditions.

In general, the pseudo outcome needs to be a quantity which represents the expected outcome of the individual taking their history up to the present as fixed, and assuming that in the future they are treated optimally. That is, in order for dWOLS to be able to estimate the optimal DTR, we need

$$E\left[\widetilde{Y}_j \middle| H_j, A_j\right] = E\left[Y^{\text{opt}} \middle| H_j, A_j\right] - \sum_{\ell=1}^{j} \mu_j(h_j, a_j).$$

The process of estimating this quantity relies on the tailoring variate, as well as the estimated parameter values.

The discussions regarding corrections in this chapter focus primarily on the application of dWOLS. However, any of the regression-type estimation procedures for DTR estimation will be subject to the same concerns when tailoring covariates are measured with error. If instead DTR estimation proceeds through the use of value search or classification-type techniques, the specific set of issues to overcome will differ.

6.2 Summary of the Proposed Methods

In this chapter we present methods for adjusting for the impacts of covariate measurement error in the estimation of an optimal dynamic treatment regime. Our proposal amounts to using regression calibration to estimate $\widehat{X}_i$, for each individual, based on replicates or a validation sample.[1] Then, we advocate for conducting an analysis using dWOLS, where $\widehat{X}_i$ is used in place of X_i within the blip, treatment-free, and treatment models directly. Doing this will produce doubly robust estimators for the blip parameters anytime that a valid pseudo outcome is constructed.

To construct a valid pseudo outcome, there are four points to consider. If you are simply wanting to estimate the effectiveness of a particular treatment, and are as such directly using the blip formulation, then replacing X_i with $\widehat{X}_i$ within the construction will consistently estimate the correct outcomes. If, instead, you are concerned with optimal DTRs and you have access to a validation sample, then the pseudo outcome can be constructed by directly modelling, in the validation sample, $E[I(H_j'\psi_j > 0)H_j'\psi_j|H_j^*]$ for each j, and using this in place of $A_j^{\text{opt}}H_j'\psi_j$ in the pseudo outcome construction. If replicate data, or other auxiliary information, is to be used in place of a validation sample, then distribu-

[1] The techniques from Chapter 3 are also appropriate.

tional assumptions may render consistent estimation of the necessary pseudo outcomes (for instance, as discussed in Theorem 6.4.3).

One final option regarding pseudo outcomes is to simply use $\widehat{X}_i$ in place of X_i and follow the standard construction. This will generally not produce consistent estimates of valid pseudo outcomes, however, we demonstrate through simulation that this simple procedure works quite well as an approximation, and is likely to be suitable for many situations, particularly where the stronger assumptions are not defensible. In practice this means that performing dWOLS with regression calibration imputed variates will produce estimators which are approximately consistent, and are doubly robust in the same way that the error-free dWOLS estimators are.

6.3 Corrections Under Classical Additive Error

For the purpose of introducing the proposed corrections, first suppose that we are considering a K stage dynamic treatment regime. We will assume that some segment of the tailoring variables are measured with additive error, and are not subject to systematic bias. That is, we observe

$$X_j^* = X_j + U_j, \tag{6.3.1}$$

for $j = 1, \ldots, K$. As before, we take A_j to denote the stage j (binary) treatment, and assume that all treatment indicators are observed without error. Moreover, we assume that $U_j \perp X_j$ and that the outcome Y is such that $E[Y|\overline{X}_K^*, \overline{A}_K, \overline{X}_K] = E[Y|\overline{X}_K, \overline{A}_K]$.[2] For notation, we will take H_j to be the history vector (at stage j) with the true covariates measured and H_j^* to be the history vector (at stage j) with the surrogate measurements. Note that, even if H_j^* contains components which are measured without error,[3] we can still take $H_j^* = H_j + U_j$. To do so, we simply set the relevant variance components of $\text{var}\,(U_j)$ to be zero.

Suppose that we denote $E[H_j|H_j^*]$ as $\widehat{H}_j$ and assume that this quantity is known.[4] Note that $\widehat{H}_j$ is a function of H_j^*. Suppose that the blip functions are all linear, so that we have

$$E[Y|H_K, A_K] = f_K(H_K) + A_K H_K' \psi_K,$$

[2]This assumption is essentially non-differential error, as it is implied by $Y \perp U_K|X_K$, though any situation where this independence does not hold but the simplification of the conditional expectation does will suffice.

[3]For instance, the previous treatments, or some additional tailoring covariates.

[4]The ensuing argument will rely only it being consistently estimable, though it is simpler to consider it a known quantity.

for some function $f_K(H_K)$ which is functionally independent of A_K. We have assumed, without loss of generality, that the blip relies on the full history vector. Any terms which need not appear in the blip function can have their corresponding coefficient set to 0.

As was argued in Section 5.5, this setting seems particularly amenable to regression calibration. At stage K, if the treatment-free and blip components are correctly specified, replacing H_K with $\widehat{H}_K$ will result in consistent parameter estimates for all of the regression parameters. However, an appealing strength of using dWOLS is that the treatment model grants the estimator double robustness. As a result, it is important to simultaneously consider the role of H_K^* in the treatment model. Theorem 6.3.1 demonstrates how, for the K-th stage of the DTR, double robustness can be restored through the use of regression calibration.

Theorem 6.3.1 (Stage K Double Robustness)**.** *The dWOLS estimator for ψ_K, obtained by performing a weighted ordinary least squares regression of Y on $\{\widehat{H}_K, A_K\widehat{H}_K\}$ with weights, $w_K(A_K, \widehat{H}_K)$ that satisfy*

$$\pi_K(\widehat{H}_K)w_K(1, \widehat{H}_K) = \left[1 - \pi_K(\widehat{H}_K)\right] w_K(0, \widehat{H}_K),$$

will be a consistent estimator so long as the blip model is correctly specified and either:

(A1) $f_K(H_K) = H_K'\beta_K$*; or*

(A2) $\pi_K(\widehat{H}_K; \widehat{\alpha}_K) \xrightarrow{p} P(A_K = 1|\widehat{H}_K)$*, as* $n \to \infty$*.*

That is to say, the estimator obtained by replacing H_K with $\widehat{H}_K$ is doubly robust.

There are two key points to note regarding this theorem. First, since we have supposed that $\widehat{H}_K$ is known, and since it is a function of H_K^*, the proof of this theorem can proceed conditioning on $\widehat{H}_K$ rather than H_K^*. This explains the rationale of using $P(A_K = 1|H_K^*)$ in the proof, which stems from conditioning on H_K^*, in place of $P(A_K = 1|\widehat{H}_K)$ in the theorem statement. In practice, $\widehat{H}_K$ will be estimated, as would $\pi_K(\cdot)$. Past simulation results suggest that using $\widehat{H}_K$ as the explanatory factor for π_K is preferable. However, the presented argument suggests that either fitted model should result in a consistent estimator, at stage K. Second, the required condition for this argument to hold was that

$$E[Y|H_K^*, A_K] = E[f_K(H_K)|H_K^*] + A_K\widehat{H}_K'\psi_K.$$

This point is important as, if we suppose that we have a pseudo outcome V, such that

$$E[V|H_j^*, A_j] = E[f_j(H_j)|H_j^*] + A_j\widehat{H}_j'\psi_j,$$

then this argument applies equally well on the weighted least squares regression of V on $\{\widehat{H}_j, A_j\widehat{H}_j\}$. We summarize this observation in Corollary 6.3.2.

Corollary 6.3.2 (Stage j Double Robustness)**.** *Suppose that V_j is such that*

$$E[V_j|H_j^*, A_j] = E[f_j(H_j)|H_j^*] + A_j\widehat{H}_j'\psi_j.$$

Then, the dWOLS estimator for ψ_j, obtained by performing a weighted ordinary least squares regression of V_j on $\{\widehat{H}_j, A_j\widehat{H}_j\}$ with weights, $w_j(A_j, \widehat{H}_j)$ that satisfy

$$\pi_j(\widehat{H}_j)w_K(1, \widehat{H}_j) = \left[1 - \pi_j(\widehat{H}_j)\right] w_j(0, \widehat{H}_j),$$

will be a consistent estimator so long as the blip model is correctly specified and either:

(A1') $f_j(H_j) = H_j'\beta_j$*; or*

(A2') $\pi_j(\widehat{H}_j) \xrightarrow{p} P(A_j = 1|\widehat{H}_j)$*, as* $n \to \infty$.

That is to say, the estimator obtained by replacing H_j with $\widehat{H}_j$ is doubly robust.

Corollary 6.3.2 provides a theoretical justification for using regression calibration to correct for the effects of measurement error in the estimation of the blip parameters. The result holds so long as the conditional mean structure is valid, and $\widehat{H}_j$ is consistent for $E[H_j|H_j^*]$. In the framing at the outset of this chapter, this result indicates how to overcome issues with the first two uses of the tailoring variables. The statement of this result indicates the necessity of deriving a process for estimating valid pseudo outcomes. To apply Corollary 6.3.2, we need to be able to consistently estimate a valid V_j.

6.4 Pseudo Outcome Estimation

In the error-free setting, we compute our pseudo outcomes as

$$\widetilde{y}_j = \widetilde{y}_{j+1} + \mu_{j+1}(h_{j+1}),$$

where $\mu_j(\cdot)$ is the j-th regret function. To estimate this quantity, we must estimate a_j^{opt}, as well as the blip parameters, ψ_{j+1}. To use a concrete example, consider a two-stage DTR, with a linear specification for the stage two blip,

$$\gamma_2(x_2, a_2) = a_2(\psi_{20} + \psi_{21}' x_2).$$

Here $a_2^{\text{opt}} = I(\psi_{20} + \psi_{21}' x_2 > 0)$. If we observe x_2 then $\widehat{a}_j^{\text{opt}} = I(\widehat{\psi}_{20} + \widehat{\psi}_{21}' x_2 > 0)$ rendering

$$\widehat{\widetilde{y}}_1 = Y^{\text{opt}} - \mu_1 - \mu_2 + \widehat{\mu}_2,$$

where $\widehat{\mu}_2 = (\widehat{a}_2^{\text{opt}} - a_2)(\widehat{\psi}_{20} + \widehat{\psi}_{21}' x_2)$. If $\widehat{\psi}_{20} = \psi_{20}$ and $\widehat{\psi}_{21} = \psi_{21}$, then $\widehat{a}_2^{\text{opt}} = a_2^{\text{opt}}$ and $\widehat{\mu}_2 = \mu_2$. This simplifies the estimated pseudo outcome to $\widehat{\widetilde{y}}_1 = Y - \mu_1$, which is the same as the theoretical quantity $\widetilde{y}_1$.

Even if $\widehat{\psi}_2 = \psi_2$, this simplification will not generally occur if x_2 is measured with error, for two reasons. First, the use of $\widehat{X}_2$ in place of X_2 will result in a residual term between the blip functions. Second, the estimated optimal treatment $\widehat{a}_2^{\text{opt}}$ may differ from the true optimal treatment. Despite this, there is a heuristic argument that suggests that forming pseudo outcomes with the regression calibration corrected covariates, $\widehat{X}$, may be a reasonable choice. Assuming that $\widehat{\psi} = \psi$, then we have

$$\widehat{\mu}_2 - \mu_2 = \left(\widehat{A}_2^{\text{opt}} - A_2^{\text{opt}}\right)\left(\psi_{20} + \psi_{21}\widehat{X}_2\right) + \left(A_2^{\text{opt}} - A_2\right)\psi_{21}'\left(X_2 - \widehat{X}_2\right). \tag{6.4.1}$$

The second term in Equation (6.4.1) has an impact dictated by $X_2 - \widehat{X}_2$. Among unbiased linear estimators of X_2, $\widehat{X}_2$ minimizes the MSE of this quantity, justifying the use of regression calibration with respect to this term. The first term relies on a difference of indicator functions. If γ_2 is significantly larger than 0 in magnitude, such that there is an unambiguous optimal treatment for the individual, then controlling $|\widehat{\gamma}_2 - \gamma_2|$ leads to $\widehat{A}_2^{\text{opt}} = A_2^{\text{opt}}$. In this situation, $\widehat{\gamma}_2$ near γ_2 will be true if $\widehat{X}_2$ is near X_2, and so we can once again rely on the justification that $\widehat{X}_2$ minimizes the MSE to motivate the selection of the regression calibration correction.

If the optimal treatment is ambiguous ($|\gamma_2| \leq \epsilon$ for a sufficiently small ϵ), then it no longer suffices to have $\widehat{\gamma}_2$ near γ_2 (as even small perturbations between these quantities may lead to $\widehat{A}_2^{\text{opt}} \neq A_2^{\text{opt}}$). However, if $\widehat{\gamma}_2$ is near γ_2, then we can also make the claim that $|\widehat{\gamma}_2|$ is relatively small. The magnitude of the first term in Equation (6.4.1) is given by $|\widehat{\gamma}_2|$, therefore, selecting an estimator to be near γ_2 will ensure that either (1) $\widehat{A}_2^{\text{opt}}$ is likely to be optimal in the event that there is a large treatment effect, or (2) that the magnitude

of the error produced will be small when $\widehat{A}_2^{\text{opt}}$ is not optimal. This provides a heuristic rationale to use the regression calibration correction to estimate the pseudo outcomes.

While this argument provides a reasonable intuition that this strategy may be effective, we can more specifically investigate the construction of pseudo outcomes through the assumptions of Corollary 6.3.2. We want to be able to form a pseudo outcome, V_j, which is valid in the sense that the conditional expectation takes the form of the conditional expectation of a treatment-free model, plus a linear blip term. The key requirement with this assumption is that, in taking on this form, the treatment-free component dictates the interpretation of the corresponding models. There are two primary interpretations for this treatment free component: either as the outcome that would be received if the individual goes on to receive treatment zero in the future (which allows for an assessment of the effect estimation), or as the outcome that would be received if the individual is treated optimally in the future (which allows for an assessment of the optimal DTR).

We consider each of these possibilities in turn, starting with an interest in effect estimation. Theorem 6.4.1 demonstrates that, without any further assumptions, pseudo outcomes can be constructed which allow for consistent estimation of the blip terms for effect estimation.

Theorem 6.4.1 (Pseudo Outcomes for Effect Estimation)**.** *Suppose that V_j is a valid pseudo outcome for effect estimation, as is described in Corollary 6.3.2. Then, if ψ_j and $\widehat{H}_j$ are known, taking $V_{j-1} = V_j - A_j \widehat{H}_j' \psi_j$ produces a pseudo outcome which is also valid for effect estimation.*

This result (alongside Corollary 6.3.2) demonstrates that, in the presence of measurement error, the effect of a treatment regime can be estimated in the data through the use of $\widehat{H}_j$ in place of H_j. In practice this means that replacing the true tailoring covariates with their estimated, regression calibration imputed values, and then conducting the standard analysis for effect estimation provides a doubly robust procedure to estimate the blip parameters. Constructing pseudo outcomes which are valid for the purpose of optimal DTR estimation is more challenging. The issue is related to the previous discussion surrounding the bias in standard pseudo outcome construction. If we consider Equation (6.4.1), then in expectation $E[X_2|X_2^*] = \widehat{X}_2$, by definition, and as a result the residual bias stems from the fact that $\widehat{A}_2^{\text{opt}}$ may not equal A_2^{opt}.

The difficulty of estimating this quantity stems from the fact that $A_K^{\text{opt}} = I(H_K' \psi_K > 0)$, which is a non-differentiable function of two unobservable quantities. Establishing an estimator, $\widehat{A}_K^{\text{opt}}$ which consistently estimates this is challenging since small perturbations of the estimated blip term lead to $|\widehat{A}_K^{\text{opt}} - A_K^{\text{opt}}| = 1$. However, if we instead focus on

correct estimation in expectation, then under certain assumptions the problem becomes tractable. First, we present Theorem 6.4.2, which provides a quantity that is theoretically computable based on observed quantities in general, before exploring how this result can be applied depending on specific assumptions.

Theorem 6.4.2 (Pseudo Outcomes for Optimal DTR Estimation)**.** *Suppose that V_j is a valid pseudo outcome for DTR estimation, as is described in Corollary 6.3.2. Then, if ψ_j is known, taking $V_{j-1} = V_j - \left(E[I(H_j'\psi_j > 0)H_j'\psi_j|H_j^*] - A_j\widehat{H}_j'\psi_j\right)$ produces a pseudo outcome which is also valid for optimal DTR estimation.*

In order to apply this result in practice, we need to devise a method to estimate

$$E[I(H_K'\psi_K > 0)H_K'\psi_K|H_K^*].$$

This is challenging to do in general, and will depend on the assumed error model and availability of auxiliary data. If a validation sample, where both H_K and H_K^* are observable for some representative subset of the population, then this quantity can be directly modelled. If no validation sample is present, then it is also possible to estimate the necessary quantity through the imposition of distributional assumptions.

Theorem 6.4.3 (Estimation under Normality of Covariates)**.** *Suppose that $U_j \sim N(\boldsymbol{0}, \Sigma)$, independently of H_j. Consider a partition of H_j^* into $\{H_j^{*,EP}, H_j^{*,EF}\}$ for the error-prone and error-free components, respectively. Further, suppose that we have $H_j^{EP} \sim N(\mu_X, \Sigma_X)$, and that $E[H_j^{EP}|H_j^*] = E[H_j^{EP}|H_j^{*,EP}]$. Then,*

$$E[A_j^{opt}H_j'\psi_j|H_j^*] = C_j\left\{1 - \Phi\left(-\frac{C_j + \dot{\mu}_j}{\dot{\sigma}_j}\right)\right\} + \dot{\mu}_j + \dot{\sigma}_j\varphi\left(-\frac{\dot{\mu}_j + C_j}{\dot{\sigma}_j}\right),$$

where $C_j = H_j^{EF\prime}\psi_j^{EF}$, $\dot{\mu}_j = \psi_j^{EF\prime}\mu_X + \psi_j^{EF\prime}\Sigma\left(\Sigma + \Sigma_X\right)^{-1}\left(H_j^{*,EP} - \mu_X\right)$, and

$$\dot{\sigma}_j^2 = \psi_j^{EF\prime}\Sigma\psi_j^{EF} - \psi_j^{EF\prime}\Sigma(\Sigma + \Sigma_X)^{-1}\Sigma\psi_j^{EF}.$$

Here, $\varphi(\cdot)$ refers to the density function for a standard normal random variable, and $\Phi(\cdot)$ the corresponding CDF.

It is worth noting that the assumptions for this result can be relaxed slightly. Error-free components which violate the assumed conditional expectation condition outlined, but which are normally distributed, can be absorbed into the error-prone component and treated as error-prone with zero error-variance. Expressing the conditions in this way was

simply more concise. The results of this theorem rely on strong assumptions regarding the unobserved, underlying variates. However, when these assumptions are satisfied, all of the necessary components can be estimated, consistently, via regression calibration. In fact, the consistency of regression calibration for the components of the distribution makes this same assumption. These results, taken in combination, provide a mechanism for estimating optimal DTRs, in a doubly robust manner, when tailoring covariates are subject to measurement error.

In light of the strength of the assumptions that are required in Theorem 6.4.3, it is sensible to consider the bias which is introduced by using $\widehat{A}_K^{\text{opt}} = I(\widehat{H}'_K\psi_K > 0)\widehat{H}'_K\psi_K$ in place of $E[I(H'_K\psi_K > 0)H'_K\psi_K|H^*_K]$. Considering the quantity

$$A_K^{\text{opt}}H'_K\psi_K - \widehat{A}_K^{\text{opt}}\widehat{H}'_K\psi_K,$$

it is clear that the residual bias in this term is driven by whether or not $\widehat{A}_K^{\text{opt}} = A_K^{\text{opt}}$. If both equal 0, then the bias is trivially 0. If both equal 1, then we can consider the expected bias (conditional on $\{H^*_K, A_K\}$), which works out as

$$E[(H'_K - \widehat{H}'_K)\psi_K|H^*_K, A_K^{\text{opt}} = 1, A_K] = 0,$$

since $\widehat{H}_K = E[H_K|H^*_K]$. If $A_K^{\text{opt}} \neq \widehat{A}_K^{\text{opt}}$ we are left with residual bias (in expectation) of the form $\pm E[H'_K\psi_K|H^*_K, A_K, A_K^{\text{opt}}]$. This parallels exactly the discussion surrounding Equation (6.4.1). If $P(A_K^{\text{opt}} \neq \widehat{A}_K^{\text{opt}}|H^*_K)$ is sufficiently small then it may be the case that the bias introduced through this process results in an outcome which differs negligibly from the true, optimal outcome. In order to assess the impact of this strategy, we consider the use of this approximate technique during the simulation studies.

6.5 Confidence Intervals

Regression calibration does not, in general, lend itself to the computation of closed-form variance estimators for the parameters of interest. Bootstrapped confidence intervals tend to be the preferred solution [7]. In the case of dWOLS, there has been little theoretical development on closed-form variance estimators. They have been derived for the single-stage setting, where the authors caution that "such variance estimates require careful calculation and coding, and so will likely not be practical for the typical analyst." and indicating that bootstrap procedures seem to perform satisfactorily in their exploratory analyses [95]. A modified bootstrap procedure, the *m-out-of-n bootstrap*, was proposed for use in Q-learning

to handle non-regularity concerns in the estimation of DTRs [12]. The proposed adaptive procedure for selecting m in Q-learning has been applied, with some success, to dWOLS [82]. It seems that, where measurement error is a concern, a bootstrap procedure would presently be most suited for estimating confidence intervals for DTR parameters.

We consider the m-out-of-n procedure, with an adaptive choice of m to construct our intervals. We outline the fundamentals of the algorithm here, though these are explored in detail by Chakraborty, Laber, and Zhao [12] and Simoneau, Moodie, Platt, and Chakraborty [82]. The method performs a standard non-parametric bootstrap, where samples of size $m < n$ are drawn (with replacement), in place of the more conventional n. We take

$$m = n^{\frac{1+\zeta(1-p)}{1+\zeta}},$$

where both p and ζ are hyperparameters, selected from the data. The parameter p is a measure of the non-regularity for the model in question, taking values in $[0, 1]$. When $p = 0$, (where we have no regularity concerns), $m = n$ and this method is equivalent to the standard bootstrap. For a fixed value of n, $m \in [n^{1/(1+\zeta)}, n]$, and so ζ can be viewed as a parameter which controls the smallest acceptable re-sample size.

We use an adaptive approach which estimates both p and ζ from our data. Consider a two-stage setting. Non-regularity concerns stem from patients for whom small perturbations in covariates lead to different optimal treatment decisions. As such, we take $\widehat{p} = \widehat{P}(\widehat{\gamma}_2 = 0)$, which we estimate as the proportion of individuals who do not admit a unique optimal treatment decision at the second stage. That is, we construct confidence sets for the second stage blip, and count the proportion of individuals for whom the range of blips computed over this set contains 0. To select $\widehat{\zeta}$, we use a double-bootstrap procedure.

We start by setting ζ to be a small value, and then draw B_1 samples of size n from the initial data. Within each of these samples, we estimate $\widehat{p}^{(b_1)}$ and the parameters of interest, $\widehat{\psi}^{(b_1)}$. We then conduct an m-out-of-n bootstrap procedure with B_2 iterations, using the current value of ζ and $\widehat{p}^{(b_1)}$ to compute $\widehat{m}^{(b_1)}$. We use these B_2 resamples to form a confidence interval around the parameters of interest. This is repeated for each of the B_1 samples. We then check the nominal coverage probability, counting the proportion of the B_1 intervals which contain the initial estimate, and if this is at the desired level, we select the present value of ζ for $\widehat{\zeta}$. Otherwise, we increment ζ and run the procedure again. The search space for ζ can be selected as necessary for the application, for instance, restricting the maximum considered value based on the smallest allowable re-sample size. Once $\widehat{\zeta}$ and $\widehat{p}$ are selected the bootstrap is performed with the estimated $\widehat{m}$.

6.6 Alternative Measurement Error Models

The previous discussion was based on the classical, additive, error model, with normally distributed error, as defined by Equation (6.3.1). However, this model was selected only for the sake of clear exposition. In Theorem 6.3.1 and Corollary 6.3.2, the proof relies solely upon the (approximately) non-differential nature of the error, and the ability to estimate $\widehat{H}_j$. Theorem 6.4.1 similarly relies upon an estimate of the conditional mean, and consistent estimation of the previous parameters, as does Theorem 6.4.2. The distributional assumptions made for the specifics of Theorem 6.4.3 are required for this result, specifically, however, similar results can be derived under any assumed error distribution.

As a result, most of this discussion could have relied on the generalized error model presented in Chapter 3. We have intentionally left the estimation of $\widehat{H}_j$ as a procedure which is independent of the DTR-specific corrections. This is generally the strategy taken with regression calibration, where the error correction takes part independently of the underlying modelling, and doing so means that these results all hold in any setting where we are able to adequately estimate the conditional mean.

The presented correction techniques lend themselves particularly well to settings where a validation sample is present. With validation data, models can be fit directly to the $E[H_j|H_j^*]$ and, once ψ_j are estimated, to $E[I(H_j'\psi_j > 0)H_j'\psi_j|H_j^*]$, using standard regression techniques. While validation data may be rare in practice, the proposed corrections allow for consistent estimation of the true blip parameter values using entirely standard software. In the event that other auxiliary data are relied upon to facilitate these corrections, particular attention will need to be paid to estimating valid pseudo outcomes.

6.7 Simulation Studies

We now demonstrate, via simulation, the potential impact of measurement error in the context of DTRs. We emphasize the issues that are present when conducting a naive analysis, and show the feasibility of regression calibration to largely correct for the errors.

6.7.1 Parameter Estimation

We begin by demonstrating the bias present in blip parameter estimates resulting from a naive analysis, and the robustness of our proposed estimation procedures. We consider a

simple one-stage setup, with $X \sim N(0,1)$, and assume that we observe two proxy measurements, given by $X_1^* \sim X + U_1$ with $U_1 \sim N(0, 0.25)$, and $X_2^* \sim X + U_2$ with $U_2 \sim t_8$. The treatment model is taken to be $P(A = 1|X_1^* = x) = \text{expit}(1 - 0.5x + 1.5\exp(x - 1))$, and the outcome model is $Y = X + \exp(X) + A(1 + X) + \epsilon$ where $\epsilon \sim N(0, 1)$, independent of all other variables. We are interested in estimating $\psi_0 = 1$ and $\psi_1 = 1$.

We consider four analyses, repeated with and without regression calibration. We fit models with (1) neither the treatment nor treatment-free models correctly specified, (2) only the treatment model correctly specified (where the treatment-free is taken to be linear), (3) only the treatment-free model correctly specified (where the treatment model is taken to be linear on the logistic scale), and (4) where both are correctly specified. We simulate 10000 data sets of size $n = 1000$. The results are summarized in Figure 6.1.

When at least one model is correctly specified (analyses (2)-(4)), the naive estimators of ψ_0 perform well. In all four scenarios the naive results are biased for ψ_1. Regression calibration results in a clear improvement over the naive estimators across analyses (2)-(4), where the bias is largely removed. There is a clear, though reduced, bias in analysis (2), where the estimates rely on the correct specification of the treatment model alone.

We extend these analyses to a variety of two-stage DTR settings, with results summarized in Appendix D, in Tables D.1-D.5. We see that whether actual treatment decisions are based on a single error-prone covariate, or on the mean of multiple proxies, the correction methods are generally applicable. The proposed corrections tend to improve estimates compared to the naive analysis, and yield results which appear broadly consistent. The corrections work well across a variety of error mechanisms. When the treatment model is badly misspecified, we see degradation in the quality of the correction.

6.7.2 Coverage Probabilities

Next, we consider three scenarios to test the applicability of the proposed bootstrap procedure. We perform the double-bootstrap procedure once under each of the scenarios, and then consider the m-out-of-n bootstrap for values of ζ surrounding the selected one. In all three scenarios we take $X_1, X_2 \sim N(0, 1)$, and observe two error-prone proxies of each measurement. A summary of the distributions of the error terms are contained in Table 6.1. For all three scenarios, we take $P(A_j = 1|X_{j1}^* = x) = \text{expit}(x)$. The outcomes for scenarios 1 and 2 are given by $Y = X_1 + X_2 + A_1(1 + X_1) + A_2(1 + X_2) + \epsilon$ where $\epsilon \sim N(0, 1)$ independent of everything else. For scenario 3, we introduce an additional binary covariate, Z_2, with $P(Z_2 = 1) = 0.5$. We then take $Y = X_1 + X_2 + A_1(1 + X_1) + A_2(1 + X_2 - Z_2 - Z_2X_2) + \epsilon$ where, again, $\epsilon \sim N(0, 1)$. Note that, if $Z_2 = 1$ then $\gamma_2 = 0$.

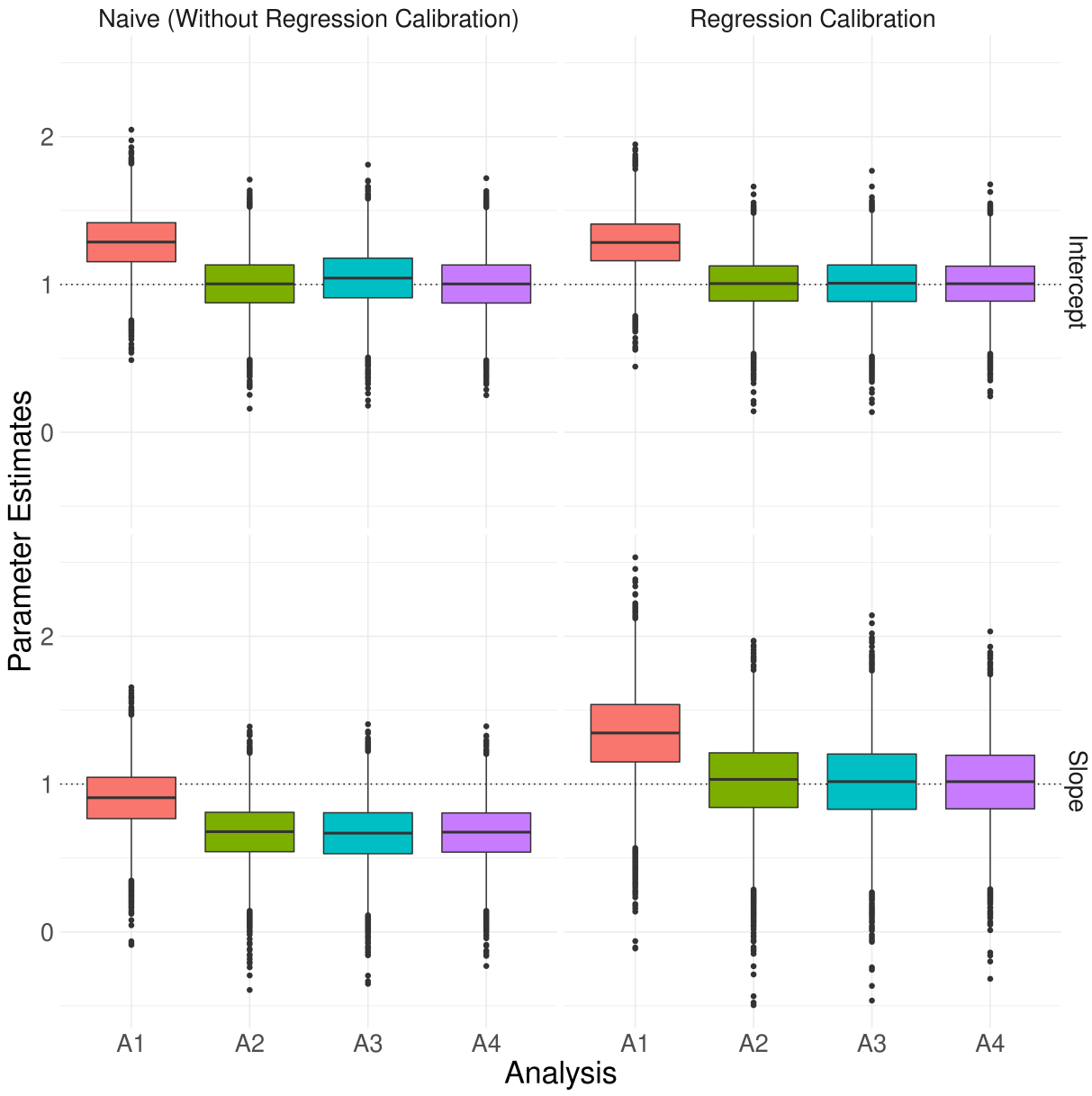

Figure 6.1: Blip parameter estimates (true values $\psi_0 = \psi_1 = 1$ indicated by dashed lines) for 10000 simulated data sets (with $n = 1000$), comparing a regression calibration corrected analysis to a naive analysis, when neither (Analysis 1), one of (Analyses 2 and 3), or both (Analysis 4) of the treatment and treatment-free models are correctly specified.

Table 6.1: Simulation study setup for the error distributions of covariates, corresponding to the first and second replicate at stages one and two. The table summarizes the distribution of the respective error terms and the type, whether additive (+) or multiplicative (×).

	Distribution				Type			
Scenario	U_{11}	U_{12}	U_{21}	U_{22}	U_{11}	U_{12}	U_{21}	U_{22}
1	$N(0,1)$	$N(0,1)$	$N(0,1)$	$N(0,1)$	+	+	+	+
2	$N(0,1)$	Unif$(-1,1)$	$N(0,1)$	Gamma$(1,1)$	+	+	+	×
3	$N(0,1)$	Gamma$(1,1)$	$N(0,0.25)$	Unif$(-1,1)$	+	×	+	+

In the first two scenarios we estimate $\widehat{\zeta} = 0.05$, while in the third scenario $\widehat{\zeta} = 0.075$. For all scenarios we form confidence intervals using (1) a traditional n-out-of-n bootstrap, (2) an m-out-of-n bootstrap where $\zeta = 0.05$ is used in the adaptive procedure, and (3) an m-out-of-n bootstrap where $\zeta = 0.10$ is used in the adaptive procedure. For the third scenario, we also include an m-out-of-n bootstrap where $\zeta = 0.075$ is used. The coverage probabilities are contained in Table 6.2. We see that the standard bootstrap procedure attained the nominal coverage in all settings. Taking the selected $\widehat{\zeta}$ met the nominal coverage levels in the second scenario, and was slightly conservative for the first and third scenarios, where taking $\zeta = 0.10$, we obtained mostly conservative intervals. In the third scenario, all procedures tended to produce conservative results.

Table 6.2: Coverage proportion for 500 repeated simulations for bootstrap coverage (2000 replicates) comparing an n-out-of-n (nn) or an m-out-of-n bootstrap based on the adaptive procedure with ζ (mn$_\zeta$). Bold values indicate those which deviate significantly from the nominal coverage of 0.95.

	Scenario One			Scenario Two			Scenario 3			
	nn	mn$_{.05}$	mn$_{.10}$	nn	mn$_{.05}$	mn$_{.10}$	nn	mn$_{.075}$	mn$_{.05}$	mn$_{.10}$
A_1	0.94	0.96	**0.97**	0.95	0.96	**0.97**	**0.97**	**0.98**	**0.98**	**0.98**
A_1X_1	0.96	**0.97**	**0.98**	0.95	0.96	0.96	**0.98**	**0.99**	**0.99**	**0.99**
A_2	0.96	**0.97**	**0.98**	0.95	0.96	**0.98**	**0.97**	**0.99**	**0.99**	**1.00**
A_2X_2	0.94	0.96	**0.97**	0.95	0.96	0.96	0.96	**0.99**	**0.98**	**0.99**
A_2Z_2	–	–	–	–	–	–	**0.97**	**0.98**	**0.98**	**0.98**
$A_2X_2Z_2$	–	–	–	–	–	–	0.97	**0.99**	**0.98**	**0.99**

6.8 Data Analysis

We now illustrate the proposed correction methods through application to STAR*D. Following Chakraborty, Laber, and Zhao [13], we model QIDS-C as a continuous covariate and consider three tailoring variables: QIDS-C measured at the start of each level (given by Q_j for stage j), the change in QIDS-C divided by the elapsed time over the previous level (referred to as QIDS slope, denoted S_j for stage j), and patient preference (denoted P_j for stage j), a binary indicator specifying whether the patient desired to switch treatment regimes ($P_j = 1$) or augment ($P_j = 0$). The outcome was taken to be $Y = -\frac{1}{2}\left(\text{QIDS-C}_1 + \text{QIDS-C}_2\right)$, where QIDS-C_j is the clinician rated QIDS score at the end of stage j.

Existing analyses of these data make the implicit assumption that clinician scores are error-free measurements. Instead, we assume that there exists a true underlying symptom score for every patient. Then both the self-assessed and the clinician scores are surrogate measures for this truth, permitting the use of generalized regression calibration. Our analysis continues to use QIDS-C as the outcome variable.

We fit the model using only the clinician ratings, only the self-reports, or using the correction where they are considered to be error-prone proxies. Following previous analyses of the data, we pose a first stage treatment model using only first stage preference (P_1) and a second stage treatment model using only second stage preference (P_2). For the first stage, the treatment-free and blip models are linear in preference (P_1), slope (S_1), and initial QIDS score (Q_1). At the second stage, the treatment-free model is linear in preference (P_2), slope (S_2), starting value (Q_2), as well as stage one treatment (A_1). The blip model used only slope (S_2) and starting value (Q_2). For each of the settings we conducted an m-out-of-n bootstrap, choosing m using the outlined adaptive procedure. Table 6.3 contains the results for parameters estimates and 95% confidence intervals.

Previous analyses have found that the only significant treatment effect was the interaction between stage one treatment and preference (A_1P_1) [12], a result that is replicated on our subset of the data when using only clinician scores. If instead we assume that the self-reported scores represent the true values, we find a significant treatment effect at stage two, with the interaction between treatment and slope (A_2S_2). However, if we perform our correction, neither of these effects remains significant, and we lack evidence for any significant treatment effects. This may be due to increased uncertainty from the error, but it nevertheless suggests further consideration is required.

Table 6.3: Results for two-stage blip coefficient estimates comparing an analysis with the regression calibration correction to naive analyses, with confidence intervals computed based on 2000 m-out-of-n adaptive bootstrap replicates. Bold values indicate treatment effects which are significant at a 95% level.

	Error Corrected	Clinician Score	Self-Reported
	Estimate (95% CI)	Estimate (95% CI)	Estimate (95% CI)
Stage One			
A_1	-0.75 (-10.04, 7.93)	-0.48 (-6.28, 5.45)	1.35 (-3.78, 6.05)
A_1P_1	2.72 (-0.19, 5.82)	**2.99 (0.90, 5.43)**	2.76 (-0.20, 5.83)
A_1Q_1	0.06 (-0.57, 0.70)	0.07 (-0.35, 0.46)	-0.09 (-0.41, 0.25)
A_1S_1	-1.54 (-6.90, 2.26)	-1.04 (-3.77, 1.08)	-0.55 (-2.31, 0.89)
Stage Two			
A_2	-0.31 (-7.05, 6.95)	1.19 (-2.88, 5.52)	-0.04 (-4.26, 4.08)
A_2Q_2	0.09 (-0.48, 0.63)	-0.02 (-0.37, 0.30)	0.08 (-0.22, 0.38)
A_2S_2	1.82 (-2.79, 4.83)	0.94 (-0.82, 2.70)	**2.74 (0.30, 5.14)**

Chapter 7

DTRs with Nonadherence

7.1 Motivation for Nonadherence Correction

Chapter 5 introduced the estimation of optimal DTRs in the error-free setting. In Chapter 6 we discussed measurement error in tailoring covariates. Until now, we have assumed that all treatment exposures are correctly measured and recorded. If this is not the case, the data are subject to nonadherence. Just as an analysis which ignored errors in tailoring covariates predictably leads to biased estimates, it is similarly intuitive that misclassified treatment indicators will result in biased estimates. However, this once again is not sufficient to conclude that error corrections are necessary. As introduced in Section 5.6, it is generally the case that an analysis which ignores the impacts of nonadherence is called an intention to treat analysis. Standard guidance advocates for the use of an ITT anytime there is nonadherence, particularly in contrast to a per-protocol (or as-treated) analysis.

There is also the consideration that the primary intention of estimating an optimal DTR may be for the assignment of future treatments. If it is likely to be the case that future individuals being prescribed treatment will have similar adherence concerns, then it may be argued that the causal effect of interest is prescription itself, rather than treatment. The intersection of these two ideas has received attention in the standard causal inference literature, though with dynamic treatment regimes, no such literature exists. The added complexity of DTRs warrants further discussion of these ideas. In brief, a naive analysis may be a sensible approach in certain scenarios where fairly restrictive assumptions are met, discussed in more detail in Section 7.3. However, we argue that generally this will not be the case. Moreover, there are settings where a naive analysis of a DTR will not produce a

causal estimate at all, owing to violations to the causal identifiability assumptions. Because of this, alternatives to intention to treat analyses are necessary.

Just as tailoring variates impacted DTRs through a variety of pathways, so too will treatment indicators. The treatment indicators serve as interaction factors in the outcome model, defining the contribution of the blip. For methods like dWOLS and G-estimation, these treatment indicators further serve as outcomes in the treatment models. We have not previously discussed errors in outcomes in this thesis. Further, past treatment indicators may serve as tailoring factors in future stages. As a result, all of the impacts discussed in Chapter 6 potentially need to be overcome. This includes requiring further considerations on the construction of pseudo outcomes. In this sense, the setting of nonadherence is more complex than the setting of errors in variables. Despite this we find that, when we can correct for the impacts of misclassification, the discrete nature of the problem renders stronger theoretical guarantees without imposing strong assumptions as compared to the errors in variables case.

We first approach nonadherence modelling based on a discussion of the likelihood, before turning towards semiparametric procedures that are more naturally motivated by the DTR setting. These corrections are presented after fully exploring the drawbacks, causally and otherwise, of a naive analysis in this setting.

7.2 Summary of the Proposed Methods

In this chapter we present methods for adjusting for the impacts of treatment misclassification in the estimation of an optimal dynamic treatment regime. The proposed method is an estimating equation approach which modifies the standard G-estimation procedure. We specify the same three models required for conducting G-estimation (the blip model, the treatment-free model, and the treatment model), where the treatment model is specified explicitly based on treatment assignment, rather than on the treatment received. In addition, we specify a model (which can be fit using validation data) which estimates the probability of the true treatment being received, given the observed factors. A similar set of estimating equations are solved as with standard G-estimation, where treatment indicators are replaced by the estimated adherence models.

The resulting estimators will be doubly robust (in the same sense that the G-estimation estimators are, with full adherence), and will be asymptotically normal under non-exceptional laws, supposing that the pseudo outcomes can be consistently estimated. The only issue with estimating pseudo outcomes stems from the situation where a previous stage's treatment is used as a tailoring variable within the blip. When this happens, adjustments (as in

Section 7.9) can be made without the need for strong assumptions to consistently estimate valid pseudo outcomes. However, just as with errors in variables, we demonstrate that these additional corrections are not completely necessary. Instead, we can follow the same process of replacing treatment indicators with their expectations, and this construction will typically suffice for valid estimation.

These proposed techniques, supposing that the models are correctly specified, can be applied whether the error-prone treatment indicator is a prescribed treatment (and therefore causes the truth) or if it is a reported treatment (and is therefore caused by the truth). This consideration is important as there is no causal interpretation for a naive analysis which is conducted when using a reported treatment, in place of the truth, supposing that the true treatment is unmeasured.

7.3 Concerns with Nonadherence

When tailoring covariates are mismeasured, a naive analysis can be justified under the idea that the mismeasured covariate can itself be a tailoring variable. If we have nonadherence, the causal effect we are estimating changes. As discussed in Sections 5.6 and 7.1, ignoring nonadherence in an analysis estimates the *intention-to-treat effect*, which is the causal effect of assigning someone treatment, rather than the effect of that treatment directly. In Figure 7.1 we show this graphically. We take A^* to represent the observable treatment, with A representing the treatment that the patient actually took. The causal effect we are interested in is the one connecting A to Y. We can see that when $A^* = A$, there is only one causal effect. When $A^* \neq A$, there is both a direct effect of A^* on Y and an indirect effect, through its impact on A. The combination of these effects constitutes the intention-to-treat effect.

The direct effect of A^* on Y may, for instance, be the result of behavioural change in patients knowing the treatment that they have been assigned to. If there is no such direct relationship we say that the *exclusion restriction* holds. Even if the exclusion restriction holds, it may not be the case that the intention-to-treat effect equals the causal treatment effect, if nonadherence is present.

In this discussion we have framed A^* as a treatment indicator which is measured prior to the true treatment. In the medical setting this may be a reasonable assumption if, for instance, A^* represents a prescribed treatment while A represents the true treatment that the individual ends up taking. However, there is a third possible treatment indicator that is worth considering: the reported treatment. If, instead of recording an individual's assigned

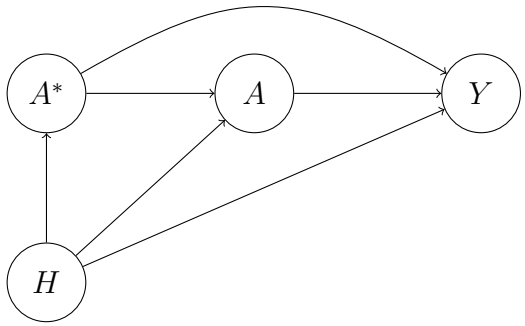

Figure 7.1: A DAG illustrating the possible change in causal interpretation when conducting an intention to treat analysis. Here, A^* is the assigned observable treatment, A is the actual treatment, Y is the outcome of interest, and H contains all possible confounders.

treatment at the time of prescription, it is reported by the individual after the fact, then the causal structure of the problem changes slightly. Suppose we label a reported treatment $A^\dagger$, then we have that A^* is an antecedent of both A and $A^\dagger$, and A is an antecedent of $A^\dagger$. This slightly more complex structure is shown in Figure 7.2.

We will refer to any (possibly misclassified) version of treatment that is recorded after actual treatment as a reported treatment, and any (possibly misclassified) version of treatment which precedes actual treatment as a prescribed treatment. In certain scenarios we may have both $\{A_j^*, A_j^\dagger\}$ recorded for all individuals. Alternatively, and more likely, only one or the other is recorded. While the scenarios are evidently related, the impact of each misclassification mechanism on the causal validity of a naive analysis differ.

If only the reported treatment is observed, then an ITT analysis cannot be conducted with a valid causal interpretation. The concern is that, as is shown in Figure 7.2, A will generally be a cause of both $A^\dagger$ and Y. If we observe $A^\dagger$ as the treatment (so that we are looking at the "causal effect of reported treatment"), and Y as the outcome, then this renders A as an unmeasured confounder, violating the causal identifiability assumptions. As a result, an ITT analysis cannot generally proceed on the basis of a reported treatment, supposing that the underlying treatment has any direct causal effect on the outcome.

A similar concern may happen in trying to conduct a standard analysis where A is observed directly, having the prescribed treatment A^* act as an unmeasured confounder. In order for this to be a concern, the prescribed treatment would need to have a direct effect on Y, not mediated through A or H. Such scenarios could plausibly exist, where, for instance, a prescribed treatment induces lifestyle changes in an individual. Even when complete adherence information is available and randomization is used, nonadherence may represent

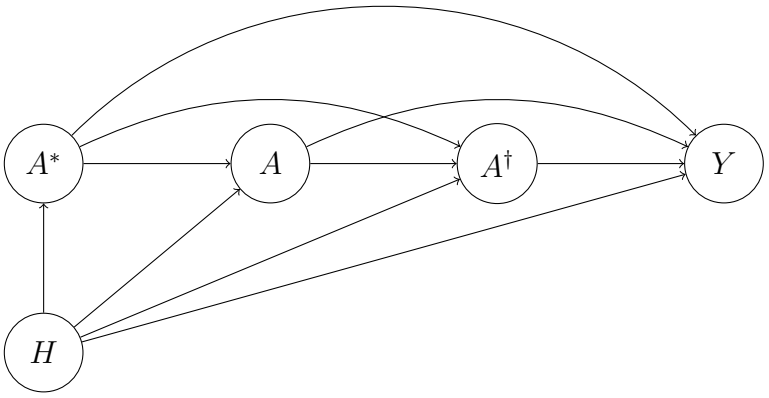

Figure 7.2: A DAG illustrating the possible change in causal interpretation when conducting an intention to treat analysis, where both treatment prescription and reported treatment are included. Here, A^* is the assigned observable treatment, A is the actual treatment, Y is the outcome of interest, and H contains all possible confounders.

a violation of the SRA [36]. Suppose that an unobserved behavioural characteristic, W_i, is such that when present ($W_i = 1$) an individual likely has a worse outcome, Y_i, regardless of treatment. If $W_i = 1$ also increases the likelihood of nonadherence, then even if $Y_i^d \perp A_{i,k}^* | H_{i,k}^*$ for all $k = 1, \ldots, K$, we will generally not have $Y_i^d \perp A_{i,k} | H_{i,k}$, because of W_i. Even when adherence information is recorded, an as-treated analysis may not be valid.

The same concerns are not present with an ITT conducted on the basis of the effect of A^*. In this case it is possible to estimate the causal effect of treatment assignment, though as the discussion in Chapter 5 alludes to, there are other possible drawbacks to estimating the ITT effect in place of directly considering intervention effectiveness. The general argument which renders an ITT useful is that, as a clinician, you do not have the ability to intervene directly on the treatment that the individual takes, you only have the ability to prescribe the treatment. As a result, the causal effect of interest is at the level of prescription. Moreover, we know that in practice, some people are going to be nonadherent outside of the confines of the study as well. By not adjusting for the adherence status we are thus estimating a more correct effect of the influence of treatment prescription on patient outcomes. While these considerations are not unreasonable, to be practically applicable they make several strong assumptions.

First, they suppose that the rates of nonadherence within the study are likely to be representative of rates of nonadherence outside of the study. While this may be the case in some designs, it is likely that the factors that influence adherence during a study and

outside of one may be different. Similarly, this line of argumentation supposes that the adherence rates are fixed and cannot themselves be influenced. If it is found that prescribing a particular treatment has a beneficial effect for patients, and it is found that there are policies which would increase adherence to this particular treatment, we cannot know from the ITT whether these policies should be pursued. This may be relevant if, for instance, a treatment that is found to be effective through an ITT is to be administered in a monitored clinical setting. The reverse is also true. It may be the case that an effective treatment had particularly bad adherence,[1] but that it is an effective treatment itself. Even from a policy perspective, there is value is disentangling the effect of prescription, and the effect of the intervention itself.

Suppose then, by way of example, we consider a single stage DTR, with a linear blip function. Further, suppose that the probability of misclassification depends only on the prescribed treatment, and no other tailoring factors. That is, assume that $A \perp X|A^*$. Assume that there is no direct effect of A^* on Y, such that $Y \perp A^*|\{X, A\}$ (called the exclusion restriction). Suppose that the outcome is truly given by

$$E[Y|X, A] = X'\beta + AX'\psi,$$

for some parameter vectors $\{\beta, \psi\}$. Straightforward calculations demonstrate that, if the regression model is fit using $\{X, A^*X\}$ in place of $\{X, AX\}$, then we will have

$$\widehat{\psi} \xrightarrow{p} \psi \left[P(A = 1|A^* = 1) - P(A = 1|A^* = 0)\right].$$

As a result, in this setting the ITT estimate of ψ will be attenuated (and as such biased) by a factor of $P(A = 1|A^* = 1) - P(A = 1|A^* = 0)$.[2] Because of this bias, it will not generally be the case that the ITT, even under the exclusion restriction, can be interpreted as the causal effect of treatment.

However, if we view the primary purpose of optimal DTR estimation as determining the optimal regime, then so long as $P(A = 1|A^* = 1) > P(A = 1|A^* = 0)$,[3] then in this setting the optimal treatment estimated by an ITT will exactly coincide with the optimal treatment. That is because, if $P(A = 1|A^* = 1) > P(A = 1|A^* = 0)$ then

$$X'\psi \geq 0 \iff X'\psi \left[P(A = 1|A^* = 1) - P(A = 1|A^* = 0)\right] \geq 0,$$

[1] Perhaps owing to something which could be changed.

[2] Note that $|P(A = 1|A^* = 1) - P(A = 1|A^* = 0)| \leq 1$. In the event that there is no misclassification this equals 1 and in the event that there is complete misclassification this equals -1.

[3] Which should be the case in most settings.

and so both parameters result in the same treatment assignment.[4] This result relies on the fact that, when the nonadherence is independent of all tailoring variables, the attenuation factor for every coefficient is exactly equal. Under these assumptions the ITT can be viewed as a valid method for estimating the optimal DTR.

The assumption that nonadherence is not related to any tailoring covariates presents a best-case scenario for the applicability of an ITT analysis. It is also an assumption that will often be violated in practice. Under even straightforward violations of this assumption it is no longer the case that an ITT will lead to the same optimal treatment rules for all individuals. It will not generally be possible to know in advance whether the estimated parameters will be attenuated or not.

There are also concerns relating to SUTVA and positivity. SUTVA requires that receiving a particular treatment results in the same potential outcome, regardless of how that treatment is administered. In a placebo-controlled drug trial, for instance, where $A_j = 0$ represents a placebo and $A_j = 1$ represents the experimental treatment, a patient who is nonadherent to $A_j = 1$ is not likely taking the placebo. Instead, a third treatment option, say $A_j = -1$ is likely required to capture the effects of nonadherence to the regime. Even when the control group is receives no treatment at all, it still may be the case that an individual who stopped taking $A_j = 1$ will have different potential outcomes as compared to the situation where they were assigned to the control group from the start.

Positivity may be unknowingly violated if, despite an acceptable treatment assignment, all individuals with a particular history are nonadherent to their assigned treatment. Suppose that, for a particular history $h_{i,j}$, $P(A_{i,j} = 1|A^*_{i,j} = 1, H_{i,j} = h_{i,j}) = 0$ and $P(A_{i,j} = 0|A^*_{i,j} = 0, H_{i,j} = h_{i,j}) = 1$, then it follows that $P(A_{i,j} = 1|H_{i,j} = h_{i,j}) = 0$ which is a positivity violation. In this setting, we would be unable to assess the efficacy of a treatment regime assigning $A_{i,j} = 1$ to an individual with history $H_{i,j} = h_{i,j}$, as it would require extrapolation beyond the available data.

One final consideration for attempting to conduct an ITT relates to the ability to correctly model the blip function. Suppose that a clinician, or other subject-matter expert, has a sense of what is important for the "true blip", perhaps through a scientific understanding of the biochemistry. There is nothing constraining the blip from an ITT to have the same form as the blip based on the recorded truth. To make matters concrete, suppose that

$$Y = f(X_1, X_2) + A_1(1 + X_1) + A_2(1 + X_2) + \epsilon,$$

[4]Note that we are not saying that the *estimated* optimal treatments will be equivalent for any given sample.

such that the blips are given by $A_1(1 + X_1)$ and $A_2(1 + X_2)$. If we try to determine what the blip would be based on A_2^* instead, this becomes complicated to do. We know that, because of the binary status of A_1^* and A_2^*, we can decompose this into some

$$Y = f^*(X_1, X_2) + A_1^* C_1(X_1) + A_2^* C_2(X_1, X_2) + \epsilon,$$

but it is not necessarily clear what these functions will be. We know that $C_2(X_1, X_2)$ is defined as $E[Y|H_2^*, A_2^* = 1] - E[Y|H_2^*, A_2^* = 0]$. Taking the form of Y based on the actual treatments, we can compute this as:

$$\begin{aligned}
&E[A_1(1 + X_1) + A_2(1 + X_2)|H_2^*, A_2^* = 1] - E[A_1(1 + X_1) + A_2(1 + X_2)|H_2^*, A_2^* = 0] \\
&= \{P(A_1 = 1|A_2^* = 1, H_2^*) - P(A_1 = 0|A_2^* = 1, H_2^*)\}(1 + X_1) \\
&\quad + \{P(A_2 = 1|A_2^* = 1, H_2^*) - P(A_2 = 0|A_2^* = 1, H_2^*)\}(1 + X_2).
\end{aligned}$$

As a result, the blip function becomes a function of the reclassification probabilities. This is potentially non-linear in X_j, which correspondingly requires different blip specifications to capture these effects. In this setting, in order to effectively fit the ITT, one must also implicitly model the misclassification process. Because of this, to do an ITT, you cannot take the same model you would have fit had you observed the truth and fit it naively. This will not produce estimates for the correct estimands.

7.4 Likelihood Based Corrections

When correcting for the effects of misclassification likelihood techniques can often be used [7]. While it is possible to estimate an optimal DTR using likelihood methods [90, 14], it is not typically done owing to the complexity of specifying the complete data generating model. Despite this, we begin by demonstrating how likelihood techniques can be used to correct for the impacts of nonadherence in DTRs. Doing this will also help to illuminate the myriad ways which nonadherence impacts the estimation of an optimal DTR.

To begin discussing nonadherence correction techniques, we will assume that we have a validation sample. We assume that either this sample is internal or that transportability assumptions hold. Define $\mathcal{M}$ and $\mathcal{V}$ to be the index sets for the main-study (ignoring all members of the validation sample), and the validation sample, respectively. We enumerate these as $1, \ldots, N_M$ and $1, \ldots, N_V$. Denote the stage j propensity score of interest to be $P(A_j = 1|H_j = h_j; \alpha_j) = \pi(h_j; \alpha_j)$, where ultimately our interest is in α_j. We model the misclassification probabilities by $P(A_j^* = a_j^*|A_j = a_j, H_j = h_j; \eta_j) = \gamma_{a_j, a_j^*}(h_j; \eta_j)$.

The setting of DTRs differs from the standard causal setting, which has been previously studied [5], since the patient history contains treatment indicators. As a result, we cannot condition on $H_j = (X_1, A_1, \ldots, X_j)$, as, in the main sample, we only have access to $H_j^* = (X_1, A_1^*, \ldots, X_j)$. This problem is circumvented entirely if we are in a setting where previous treatment does not influence future treatment, except through other covariates. If we are able to assume that $A_j \perp A_k | \overline{X}_j$, for all $k < j$, then it is the case that

$$P(A_j = 1|H_j = (\overline{X}_j, \overline{A}_{j-1})) = P(A_j = 1|\overline{X}_j) = P(A_j = 1|H_j^* = (\overline{X}_j, \overline{A}_{j-1}^*)).$$

This *longitudinal treatment independence assumption* (LTIA) is theoretically verifiable, given a validation sample, since it equates to a hypothesis that some subset of the α_j parameters are simultaneously 0. If this assumption does not hold then we need to take into account the covariate error during correction.

Similar considerations need to be given to the misclassification models. In order to model $\gamma_{a_j,a_j^*}(h_j;\eta_j)$ we have assumed that the relevant functional form depends on H_j. To overcome this, we could make an assumption, either through independence or functional form, that $\gamma_{a_j,a_j^*}(h_j;\eta_j)$ does not depend on $\overline{A}_{j-1}$. We could, for instance, take misclassification probabilities to depend only on the assigned treatment at that time. That is, misclassification probabilities at each stage are constant. This assumption could be relaxed, allowing the misclassification probabilities to vary based on $\overline{X}_j$. Alternatively, noting that η_j (and $\gamma(\cdot)$ more generally) represent nuisance parameters in the model, we can introduce additional nuisance parameters and model $P(A_j^* = a_j^*|A_j = a_j, H_j^*; \eta_j^*) = \gamma^*_{a_j,a_j^*}(h_j^*;\eta_j^*)$. We will assume that we specify this model as well. If conditioning on H_j is equivalent to conditioning on H_j^* then $\eta_j = \eta_j^*$.

We now derive the full likelihood of the treatment parameters, with and without the LTIA. After introducing an approach based on the complete likelihood, we discuss ways that this may be pursued in a computationally feasible manner, before turning to corrections applied to more common DTR estimation techniques.

7.4.1 Full Likelihood Corrections

If we make the LTIA, then we can write the joint likelihood of $(\alpha_j, \eta_j, \eta_j^*)$, as

$$L(\alpha_j, \eta_j, \eta_j^*) = \left\{ \prod_{i=1}^{N_M} \left(\pi(h_{ij}^*; \alpha_j) \gamma_{1,a_{ij}^*}^*(h_{ij}^*; \eta_j^*) + (1 - \pi(h_{ij}^*; \alpha_j)) \gamma_{0,a_{ij}^*}^*(h_{ij}^*; \eta_j^*) \right) \right\} \times \prod_{i=1}^{N_V} \pi(h_{ij}; \alpha_j)^{a_{ij}} (1 - \pi(h_{ij}; \alpha_j))^{1-a_{ij}} \gamma_{a_{ij},a_{ij}^*}(h_j; \eta_j).$$

The likelihood expresses the joint conditional density of (A_j, A_j^*) given H_j over the validation sample. Braun et al. [5] model only the conditional density of A_j given H_j, making the second term in this expression the product over $\pi(h_{ij}; \alpha_j)$. Doing this would mean that we do not require estimation of η_j, though it ignores some of the recorded data. To get an estimate for α_j we can directly optimize this joint likelihood, or we could estimate η_j^* in the validation sample and then use $\widehat{\eta}_j^*$ in the likelihood estimation.

In order to relax the LTIA, we will quickly find that using the main sample with a likelihood approach is difficult. The reason is that we need to sum over the 2^j possible $\overline{A}_j$ vectors, and attempting to further break down this expression will require us to condition on future information. This is not a modelling setup which will be advantageous to pursue. If we make the assumption that $P(A_j = 1 | H_j, \overline{A}_{j-1}^*) = P(A_j = 1 | H_j; \alpha_j)$, that is that past treatment assignment is conditionally independent of future actual treatments, given the observed covariates, and we are willing to pose a direct model, specifying misclassification probabilities, as $P(\overline{A}_{j-1} | \overline{X}_j, \overline{A}_{j-1}^*)$, then we can write the full likelihood as

$$L(\alpha_j, \eta_j, \cdot) = \left\{ \prod_{i=1}^{N_M} \left(\sum_{\overline{a}_j} P(A_j = a_j | \overline{X}_{ij}, \overline{A}_{j-1} = \overline{a}_{j-1}; \alpha_j) P(\overline{A}_{j-1} | \overline{X}_{ij}, \overline{A}_{i,j-1}^*) \right) \right\} \times \prod_{i=1}^{N_V} P(A_{ij} | H_{ij}; \alpha_j) P(A_{ij}^* | A_{ij}, H_{ij}; \eta_j).$$

There are plausible scenarios where this model could be used. For instance, when misclassification depends only on the assigned treatment.[5]

These techniques can all be used to derive estimates for α_j. This gives us the capacity to use weighting techniques based on propensity scoring, whether that is in the place of

[5] That is, A_j^* uniquely specifies the probability of misclassification at stage j.

treatment models, as are required for dWOLS or G-estimation, or for use in a weighting or subclassification scheme used to correct for the error present in those models. These parameter estimates will allow for an estimation of $E[A_j|A_j^*, H_j]$, which is a term relevant for the outcome models in the regression based procedures.

In place of using these parameter estimates in these existing optimal DTR estimation techniques, an alternative strategy would be to work through the likelihood of the complete data trajectory. This requires the specification of a conditional model for X_j. We assume that either $P(A_j|H_j, \overline{A}_{j-1}^*) = P(A_j|H_j; \alpha_j)$, or that H_j contains $\overline{A}_{j-1}^*$, where assignment is meaningful. For convenience, take H_0, A_0, A_0^* to all be constants. Then,

$$\begin{aligned}
L &= \left\{ \prod_{i=1}^{N_V} \prod_{j=1}^{K} f(X_{ij}|H_{i,j-1}, \overline{A}_{i,j-1}^*) P(A_{ij}|H_{ij}; \alpha_j) P(A_{ij}^*|A_{ij}, H_{ij}; \eta_j) \right\} \\
&\times \left\{ \prod_{i=1}^{N_M} \prod_{j=1}^{K} \sum_{\overline{a}_j} f(X_{ij}|\overline{X}_{i,j-1}, \overline{a}_{j-1}, \overline{A}_{i,j-1}^*) P(A_j = a_j|\overline{X}_{ij}, \overline{a}_{j-1}; \alpha_j) P(A_{ij}^*|\overline{A}_j = \overline{a}_j, \overline{X}_{ij}; \eta_j) \right\} \\
&= \prod_{j=1}^{K} \left\{ \prod_{i=1}^{N_V} f(X_{ij}|H_{i,j-1}, \overline{A}_{i,j-1}^*) P(A_{ij}|H_{ij}; \alpha_j) P(A_{ij}^*|A_{ij}, H_{ij}; \eta_j) \right. \\
&\left. \times \prod_{i=1}^{N_M} \sum_{\overline{a}_j} f(X_{ij}|\overline{X}_{i,j-1}, \overline{a}_{j-1}, \overline{A}_{i,j-1}^*) P(A_j = a_j|\overline{X}_{ij}, \overline{a}_{j-1}; \alpha_j) P(A_{ij}^*|\overline{A}_j = \overline{a}_j, \overline{X}_{ij}; \eta_j) \right\}.
\end{aligned} \tag{7.4.1}$$

This is related to the idea of the *g-computation algorithm* [68]. However, the likelihood of the complete data trajectory will generally be computationally insurmountable.

7.4.2 Semiparametric Approach to the Likelihood

As a natural extension of this likelihood estimation procedure, we can use kernel density estimation, to make the model semiparametric. If we drop the $\gamma_{a_j, a_j^*}(H_j; \eta_j)$ term from the above likelihood, and note that

$$\gamma_{a_j, a_j^*}^*(h_j^*; \eta_j^*) = \frac{f_{A_j, A_j^*, H_j^*}(a_j, a_j^*, h_j^*)}{f_{A_j, H_j^*}(a_j, h_j^*)},$$

then, the two joint densities in the validation sample can be estimated with KDE. This gives $\widehat{f}(a_j, a_j^*, h_j^*)$ and $\widehat{f}(a_j, h_j^*)$, which we can use to replace $\gamma^*_{a_j,a_j^*}(h_j^*;\eta_j^*)$ as $\widehat{\gamma}^*_{a_j,a_j^*}(h_j^*;\eta_j^*)$. Doing so then allows us to define

$$\begin{aligned}\widehat{L}(\alpha_j) = &\left\{\prod_{i=1}^{N_M}\left(\pi(h_{ij}^*;\alpha_j)\widehat{\gamma}^*_{1,a_{ij}^*}(h_{ij}^*;\eta_j^*) + (1-\pi(h_{ij}^*;\alpha_j))\widehat{\gamma}^*_{0,a_{ij}^*}(h_{ij}^*;\eta_j^*)\right)\right\}\\ &\times\prod_{i=1}^{N_V}\pi(h_{ij};\alpha_j)^{a_{ij}}(1-\pi(h_{ij};\alpha_j))^{1-a_{ij}},\end{aligned}$$

which can be solved to estimate α_j. When not making the LTIA, we can replace the models for $f(X_j|H_{j-1},\overline{A}^*_{j-1})$ and $P(A_j^*|A_j,H_j;\eta_j)$ with kernel density estimates. The expression in this case can be simplified by limiting the contribution in the validation sample. To see this, consider that α_j is only contained in the j-th product term in Equation (7.4.1), so

$$\begin{aligned}L(\alpha_j,\cdot) = &\prod_{i=1}^{N_V} f(X_{ij}|H_{i,j-1},\overline{A}^*_{i,j-1})P(A_{ij}|H_{ij};\alpha_j)P(A_{ij}^*|A_{ij},H_{ij};\eta_j)\\ &\times\prod_{i=1}^{N_M}\sum_{\overline{a}_j} f(X_{ij}|\overline{X}_{i,j-1},\overline{a}_{j-1},\overline{A}^*_{i,j-1})P(A_j=a_j|\overline{X}_{ij},\overline{a}_{j-1};\alpha_j)P(A_{ij}^*|\overline{A}_j=\overline{a}_j,\overline{X}_{ij};\eta_j).\end{aligned}$$

In this formulation, it is far more evident that we may wish to use

$$\begin{aligned}L(\alpha_j,\cdot) = &\prod_{i=1}^{N_V}\pi(h_{ij};\alpha_j)^{a_{ij}}(1-\pi(h_{ij};\alpha_j))^{1-a_{ij}}\\ &\times\prod_{i=1}^{N_M}\sum_{\overline{a}_j} f(X_{ij}|\overline{X}_{i,j-1},\overline{a}_{j-1},\overline{A}^*_{i,j-1})P(A_j=a_j|\overline{X}_{ij},\overline{a}_{j-1};\alpha_j)P(A_{ij}^*|\overline{A}_j=\overline{a}_j,\overline{X}_{ij};\eta_j),\end{aligned}$$

where we can once again replace the nuisance models with kernel density estimates. These techniques could once again be applied to derive estimates of the nuisance parameters for use in alternative DTR estimation techniques, or alternatively, through an application of g-computation. These techniques for estimating the parameters of the true treatment assignment model take inspiration from Braun et al. [5], and demonstrate the complexity inherent to the setting of nonadherence. However, taking the likelihood framing explicitly is not the most natural approach to DTR estimation.

In particular, dWOLS and G-estimation are both techniques that rely primarily on the

theory of M-estimation rather than on likelihood theory directly. As a result, corrections for the impacts of nonadherence which are based in this semiparametric framework may be more natural to explore. If a validation sample is present, then the actual misclassification model can be estimated using standard modelling techniques, which may be likelihood or quasi-likelihood based, but which are likely more familiar, and computationally simple, for practitioners to leverage. Owing to this, we turn towards considering how corrections to estimation techniques for optimal DTRs can be applied, directly.

7.5 Modified G-Estimation

We first present the complete modification to the process of G-estimation, under the assumption of patient nonadherence, and then demonstrate the consistency of this estimator. For $j = 1, \ldots, K$ define

$$\begin{aligned} \pi_j^*(H_{i,j}^*, A_{i,j}^*) &= P(A_{i,j} = 1 | H_{i,j}^*, A_{i,j}^*); \\ \nu_j^*(H_{i,j}^*) &= E[\nu_j(H_{i,j}) | H_{i,j}^*, A_{i,j}^*]; \\ C_j^*(H_{i,j}^*) &= E[C_j(H_{i,j}) | H_{i,j}^*, A_{i,j}^*]. \end{aligned}$$

Then, take $\widetilde{V}_{i,K+1} = Y_i$ and, for all $1 \leq j \leq K$, define

$$\widetilde{V}_{i,j} = \widetilde{V}_{i,j+1} + \left[A_{i,j}^{\text{opt}} - \pi_j^*(H_{i,j}^*)\right] C_j^*(H_{i,j}^*). \tag{7.5.1}$$

With these quantities defined, we take U_j^* to be given by the set of equations

$$U_j^* = \sum_{i=1}^{n} \lambda_j^*(H_{i,j}^*) \left\{A_{i,j}^* - P(A_{i,j}^* = 1 | H_{i,j}^*)\right\} \left\{\widetilde{V}_{i,j+1} - \pi_j^*(H_{i,j}^*, A_{i,j}^*) C_j^*(H_{i,j}^*; \psi_j) + \theta_j^*(H_{i,j}^*)\right\}. \tag{7.5.2}$$

These correspond to Equation (5.4.4) for the standard G-estimation procedure (that is, when there is complete adherence). Estimators for the parameters of interest can be derived by solving $U_j^* = 0$. In Theorem 7.5.1 we demonstrate that this estimation procedure will result in consistent estimators for the blip parameters.

Theorem 7.5.1. *Suppose that for $j = 1, \ldots, K$ and $i = 1, \ldots, n$, we know $P(A_{i,j}^* | H_{i,j}^*)$ and $\pi_j^*(H_{i,j}^*, A_{i,j}^*)$, and we correctly specify the form of $C_j^*(H_{i,j}^*; \psi_j)$. Then the $\widehat{\psi}_j$ which are estimated by solving $U_j^*(\widehat{\psi}_j) = 0$ are consistent for the true ψ_j, under the following independence assumptions (I.A.):*

I.A. (1): $E[V_{j+1}(H_j)|H_j, A_j, \overline{A}_j^*] = E[V_{j+1}(H_j)|H_j, A_j]$ *for all* $j = 1, \ldots, K$.

I.A. (2): $E[C_j(H_j)|A_j = 1, H_j^*, A_j^*] = E[C_j(H_j)|H_j^*, A_j^*]$ *for all* $j = 1, \ldots, K$.

I.A. (3): $E[\nu_j(H_j)|H_j^*, A_j^*] = E[\nu_j(H_j)|H_j^*]$ *for all* $j = 1, \ldots, K$.

In order to assess the viability of this strategy, it is worth considering how reasonable the independence assumptions are. The first assumption requires that there is no predictive information contained in the treatment assignment, supposing that we know the true treatment and history, for each individual. This assumption can be viewed, in a sense, as an extension of the SUTVA. If a patient receives $A_j = 1$ we are claiming that it ought not matter whether the person had originally been prescribed $A_j^* = 1$, so long as we have the complete relevant history.

The second independence assumption is, at face value, stronger. It states that there is no mean difference in the contrast between those who actually take the treatment at time j ($A_j = 1$) and those who do not ($A_j = 0$), given the observed history and treatment assignment. This will be violated in the event that, for instance, previous compliance is related to current compliance. Fortunately, this assumption can be discarded entirely if the analyst is instead willing to specify the model $E[C_j(H_j; \psi_j)|A_j = 1, H_j^*, A_j^*]$. We would simply replace $C_j^*(H_j^*)$ with this definition, and the proof proceeds as written. This is a more challenging quantity to model, generally, but it provides a mechanism for circumventing the need for this independence assumption. We proceed assuming that either this assumption is reasonable, or else this model can be specified directly.

The final assumption requires that the treatment-free component of the stage j Q-function, $\nu_j(H_j)$, is not predicted by treatment assignment at stage j, given the history up to stage j. This is a seemingly reasonable assumption to make. In the event of complete adherence, $\nu_j(H_j)$ is functionally independent of A_j by definition, and so conceivably should be independent of treatment assignment as well. This may not be the case if, for instance, past adherence status is used to inform treatment in the present but is not recorded in the available data. However, if there are factors being used to inform treatment assignment, which are not being collected, there are likely violations to the SRA, which would mean that no causal analysis can proceed.

Just as with G-estimation under complete adherence, it will not often be the case that $P(A_{i,j}^*|H_{i,j}^*)$ or $\pi_j^*(H_{i,j}^*, A_{i,j}^*)$ are known. However, if we can model and estimate these from data then by "stacking" estimating equations the same consistency result holds. Under consistent estimation of these parameters, the results from Theorem 7.5.1 will remain. This leaves the need to specify $\lambda_j^*(H_j^*)$ and $\theta_j^*(H_j^*)$. In the case of complete adherence, $\theta_j(H_j)$

was used to endow G-estimation with the property of double robustness, and we can do the same here. If we take $\theta_j^*(H_j^*) = -\nu_j^*(H_j^*)$. Supposing that $\pi_j^*(H_j^*, A_j^*)$ and $C_j^*(H_j^*; \psi_j)$ are correctly specified, then if either $P(A_j^* = 1|H_j^*)$ or $\theta_j^*(H_j^*)$ were to be correctly specified, the resulting estimating equations are unbiased. It is worth noting that the estimating equations require $\pi_j^*(H_j^*, A_j^*)$ and $C_j^*(H_j^*)$ to be correctly specified. As a result, if there is strong evidence regarding the distribution $P(A_j|H_j^*)$, then the specification of $\pi_j^*(H_j^*, A_j^*)$ implies the correct specification of the treatment model. Similar considerations may happen with components of $C_j^*(H_j^*)$, which may rely on treatment indicators from previous stages.

Finally, the specification of $\lambda_j^*(H_j^*)$, as with $\lambda_j(H_j)$, can be arbitrary so long as it depends solely on H_j^* and is the same dimension as the parameter ψ_j. If we consider the class of estimating equations characterized by arbitrary $\lambda_j^*(\cdot)$, and write $U_j^* = \sum_{i=1}^n \lambda_j^*(H_{i,j}^*)\widetilde{U}_{i,j}$, then Morton [58] demonstrates that the optimal choice for $\lambda_j^*(H_j^*)$ is given by

$$\lambda_j^*(H_j^*) = E\left\{ \frac{\partial}{\partial \psi_j}\widetilde{U}_j \middle| H_j^* \right\} E\left\{ \widetilde{U}_j^2 \middle| H_j^* \right\}^{-1}.$$

This choice for $\lambda_j^*(H_j^*)$ is analogous to that derived by Robins [70] when discussing locally efficient estimators under complete adherence. The first term in this expression simplifies to $\frac{\partial}{\partial \psi_j} C_j^*(H_j^*; \psi_j)$, while the second is generally quite complex. In the event that all of the models are correctly specified and $\text{var}(\widetilde{V}_j|H_j^*, A_j^*) = \text{var}(\widetilde{V}_j|H_j^*)$, then this term will simplify to $\text{var}(\widetilde{V}_j|H_j^*)\text{var}(A_j^*|H_j^*)$. Assuming that these terms are constant, they can be dropped from the estimating function entirely, and we will have $\frac{\partial}{\partial \psi_j} C_j^*(H_j^*; \psi_j)$ as the optimal choice of $\lambda_j^*(H_j^*)$. We will take $\lambda_j^*(H_j^*) = \frac{\partial}{\partial \psi_j} C_j^*(H_j^*; \psi_j)$, and note that in any specific implementation, the optimal choice may be worked out.

7.6 Modelling Nonadherence

We have made use of models which describe the propensity of a patient to have taken treatment, conditional on the observed information. We have explicitly used these models as $\pi_j^*(H_j^*, A_j^*)$, and have implicitly made use of them in the models for $C_j^*(H_j^*) = E[C_j(H_j)|H_j^*, A_j^*, A_j = 1]$. The ability to adequately correct for the effects of nonadherence in this analysis depends on the ability to model this process of nonadherence, reliably. If these models are known explicitly, can be estimated from data, or can be reasonably specified based on subject-matter expertise, then the modified G-estimation procedure can proceed. The previous discussion on likelihood estimation provides one such mechanism for

modelling these parameters. In practice, the use of logistic regression models (or similar) is likely more accessible.

The estimation procedure is designed to allow for posited models on patient adherence to be used in place of actual or estimated models. This allows for a sensitivity analysis to be performed, based on the degree of nonadherence present in the data, whenever the requisite auxiliary data or model estimates are unavailable. In the unlikely event that the treatment probabilities are known precisely, these can be used exactly as described.

In the event that there is a validation sample such that, for every time $j = 1, \ldots, K$ there is some subset of individuals $i = 1, \ldots, n'_j$ with both $A^*_{i,j}$ and $A_{i,j}$ measured, then models for the required components can be fit using standard estimation techniques. One possible technique for doing so would be leveraging the likelihood results that were presented in the previous sections. This provides an approach, motivated directly from common DTR estimation techniques, that demonstrates the possible utility of the previous discussion. Oftentimes, however, it may be more natural to use alternative modelling techniques (say via generalized estimating equations, or generalized linear models more specifically). The same general method can be applied when the validation data are from an external sample. In this case, the summation will run from $i = 1, \ldots, n_1, n_1 + 1, \ldots, n_1 + n_2$ where n_1 is the size of the regular sample and n_2 is the size of the external validation sample. All of these corresponding estimators are based on M-estimation theory, and the corresponding estimating equations can be stacked onto the estimating equation U^*_j in the same manner that the treatment assignment and treatment-free models are.

If, in place of a validation sample other auxiliary data are available, it is sometimes still possible to consistently estimate the required misclassification probabilities. Buonaccorsi [6], in Section 2.6, discusses how likelihood techniques can be exploited to estimate misclassification rates, and the true marginal probability, based replicated measurements. These results are summarized thoroughly by Walter [97]. As expressed they demonstrate that it is often possible to estimate $P(A_j = 1)$ and $P(A^*_j = 1|A_j)$ from data which has three or more replicated values; the drawback to a direct application of these techniques is that we would need misclassification to be independent of any tailoring variates, at which point the need for a correction has been called into question.[6] Still, it may be possible, depending on the available data, to use a similar decomposition to approximate the necessary models.

[6]Such a technique may still be useful when using $A^\dagger_j$ rather than A^*_j, as an ITT is never valid based off of $A^\dagger_j$.

Note that, for $\pi_j^*(H_j^*, A_j^*)$, we can write

$$\pi_j^*(H_j^*, A_j^*) = P(A_j = 1|H_j^*, A_j^*) = \frac{P(A_j = 1, A_j^*|H_j^*)}{P(A_j^*|H_j^*)} = \frac{P(A_j^*|A_j = 1, H_j^*)P(A_j = 1|H_j^*)}{P(A_j^*|H_j^*)}.$$

Considering a model for $P(A_j^* = a|H_j^*)$, we can write

$$\begin{aligned} &P(A_j^* = a|H_j^*) \\ &= P(A_j^* = a|H_j^*, A_j = 1)P(A_j = 1|H_j^*) + P(A_j^* = a|H_j^*, A_j = 0)P(A_j = 0|H_j^*). \end{aligned}$$

Now, because A_j^* is assumed to be observed, we can write down the model likelihood for $P(A_j^* = 1|H_j^*)$, broadly, as

$$L = \prod_{i=1}^{n} P(A_j^* = 1|H_{i,j}^*)^{A_{i,j}^*} \left(1 - P(A_j^* = 1|H_{i,j}^*)\right)^{1-A_{i,j}^*}.$$

Suppose that we are able to specify a parametric model for each of $P(A_j^* = 1|H_j^*, A_j = 1)$, $P(A_j^* = 1|H_j^*, A_j = 0)$, and $P(A_j = 1|H_j^*)$. Suppose that these three models are given by $f_{1|1}^{(j)}(H_j^*; \gamma_{1|1}^{(j)})$, $f_{1|0}^{(j)}(H_j^*; \gamma_{1|0}^{(j)})$, and $f_1^{(j)}(H_j^*; \gamma_1^{(j)})$, respectively. By using the breakdown given for $P(A_j^* = a|H_j^*)$, with these three models on hand, we can write down the likelihood expression for $\{\gamma_{1|1}^{(j)}, \gamma_{1|0}^{(j)}, \gamma_1^{(j)}\}$ as

$$\begin{aligned} L(\gamma_{1|1}^{(j)}, \gamma_{1|0}^{(j)}, \gamma_1^{(j)}) = \prod_{i=1}^{n} &\left\{ f_{1|1}^{(j)}(H_{i,j}^*) f_1^{(j)}(H_{i,j}^*) + f_{1|0}^{(j)}(H_{i,j}^*) \left(1 - f_1^{(j)}(H_{i,j}^*)\right) \right\}^{A_{i,j}^*} \\ &\times \left\{ 1 - f_{1|1}^{(j)}(H_{i,j}^*) f_1^{(j)}(H_{i,j}^*) - f_{1|0}^{(j)}(H_{i,j}^*) \left(1 - f_1^{(j)}(H_{i,j}^*)\right) \right\}^{1-A_{i,j}^*}. \end{aligned}$$

In this expression, to simplify notation, we have suppressed the dependence of $f_\cdot^{(j)}$ on $\gamma_\cdot$. In general there will be identifiability concerns with directly optimizing this model, and as a result, more data or further assumptions will be required to estimate these parameters. Suppose that for each individual, k_i repeated measurements of A_j^* are available. Further, suppose that, given $\{A_j, H_j^*\}$, these replicates are independent of one another. In this setting, by expanding the notation so that $f_{1|a}^{(j,\ell)}$ corresponds to $P(A_{j,\ell}^* = 1|H_j^*, A_j = a)$,

where $A^*_{j,\ell}$ is the ℓ-th replicate, then

$$L(\gamma^{(j,\cdot)}_{1|1}, \gamma^{(j,\cdot)}_{1|0}, \gamma^{(j)}_1) = \prod_{i=1}^{n} \prod_{\ell=1}^{k_i} \left\{ f^{(j,\ell)}_{1|1}(H^*_{i,j}) f^{(j)}_1(H^*_{i,j}) + f^{(j,\ell)}_{1|0}(H^*_{i,j}) \left(1 - f^{(j)}_1(H^*_{i,j})\right) \right\}^{A^*_{i,j,\ell}}$$
$$\times \left\{ 1 - f^{(j,\ell)}_{1|1}(H^*_{i,j}) f^{(j)}_1(H^*_{i,j}) - f^{(j,\ell)}_{1|0}(H^*_{i,j}) \left(1 - f^{(j)}_1(H^*_{i,j})\right) \right\}^{1-A^*_{i,j,\ell}}.$$

Depending on the assumptions made, it may be the case that some parameters can be shared between replicates, potentially simplifying the assumptions here. Still, given any particular model form, a sufficient number of replicates (and sample size), and possibly misclassification assumptions, this decomposition of the likelihood can be used to derive estimates for the parameters required to compute $\pi^*_j(H^*_j, A^*_j)$. Despite the mathematical feasibility of such a possibility, it seems to us that the use of replication data for nonadherence specifically may be less fruitful than the use of replication data for correcting for the impacts of errors in variables. This is because in order for these techniques to apply, multiple, conditionally independent, potentially misclassified treatment indicators need to be available. This strategy is certainly useful when, for instance, the binary indicator of note is disease status which is the result of a (possibly faulty) test. In this setting it is possible to take several tests, each of which reports a possibly faulty result, but taken together they can be viewed as replicate measures of the binary response. In the event that the binary response represents a treatment prescription, however, most settings would appear to not have an obvious analogue. We present this possibility for parameter estimation in the event that a specific application of DTRs have data which correspond to this structure, but do not otherwise pursue these estimators in depth.

If no auxiliary data are available, but model estimates are available from existing literature, then these can be used as though they were truth, adjusting the standard errors based on the particular form. This can be viewed as a special case of the external validation sample, and will be subject to the same asymptotic variances. If there are no existing estimates, and no auxiliary data, then we recommend using these estimators to conduct a sensitivity analysis. Suppose that, despite a lack of auxiliary data, or pre-existing estimates of the rates of misclassification, subject-matter experts are able to make educated guesses at the magnitude of the impact different covariates have on the rates of nonadherence. Perhaps these are informed through studies on similar treatment regimes, or anecdotally from clinical practice. If a model for $\pi^*_j(H^*_j, A^*_j)$ can be specified, then by filling in the relevant parameters as though they were the truth, the proposed method will provide consistent estimates of the blip under these assumptions. A sensitivity analysis proceeds by considering several possibilities for the true parameters, along a range of plausible values,

and determining the impact on the estimated parameters.

Specifically, one may take

$$\pi_j^*(H_j^*, A_j^*) = \text{expit}\left(\begin{bmatrix} H_j^{*\prime} \\ A_j^* \end{bmatrix} \alpha_j^*\right).$$

Then, varying α_j^* over a pre-specified grid leads to plausible misclassification models and the modified G-estimation procedure can proceed assuming each is truth. In order for this to be applied, as specified, and for the estimators to remain computationally feasible, H_j^* must not be *too* high-dimensional.[7] This approach leads to a set of estimated parameter values for each ψ_j, which captures the impact of nonadherence on the parameter estimates. Further, the estimated optimal treatment can be computed for each individual, which allows the analyst to determine, for any specific person, what would need to be true for A_j^{opt} to be 1 or 0. In this sense it is possible to show for which individuals the presence of nonadherence is likely to alter the estimated optimal treatment assignment, and for which the optimal assignment is fairly resilient to nonadherence.

7.7 Asymptotic Distribution and Inference

At stage K, if we parameterize $\theta_K^*(\cdot)$ with β_K, $\pi_K^*(\cdot)$ with α_K, $P(A_K = 1|H_K^*)$ with γ_K, $\widetilde{H}_K(\cdot)$ (the history vector with treatment indicators replaced as described) with ζ_K, and $C_K^*(\cdot)$ with ψ_K, then if all models are selected such that these estimators can be framed as the solution to unbiased estimating equations, the full estimation procedure depends on the stacked version of these estimators. In general, we will have that

$$U_K^* = \begin{bmatrix} U_{\text{Treatment Indicator}}(\zeta_K) \\ U_{\text{Treatment Free}}(\beta_K) \\ U_{\text{Treatment}}(\alpha_K) \\ U_{\text{Treatment Assignment}}(\gamma_K) \\ U_{\text{G-Estimation}}(\beta_K, \alpha_K, \gamma_K, \zeta_K, \psi_K) \end{bmatrix}.$$

Note that while some of these components will be independent by assumption, it is possible for there to be a reliance between the treatment indicator and the treatment, treatment assignment, or treatment free components of the estimating equations.

For each $j < K$, similar sets of estimating equations can be expressed. Note that

[7] Or else, certain values of α_j^* must be restricted, to limit the size of the grid considered.

for each $j < K$, in place of Y we use $\widetilde{V}_j$, which depends explicitly on ψ_{j+1}, α_{j+1}, and ζ_{j+1}. Then, in addition to requiring the components at stage j, the estimators rely on the parameters from previous stages. The same form of dependence will occur through the use of the ζ_j parameters, wherein they may be required as tailoring factors or predictors in other estimating equations. Define U_j^* for $j = 1, \ldots, K-1$ in a similar way. Then, the complete stacked estimating equation is taken to be

$$U^* = \begin{bmatrix} U_K^* \\ U_{K-1}^* \\ \vdots \\ U_2^* \\ U_1^* \end{bmatrix}. \tag{7.7.1}$$

In this framing, whether we take the required parameters to be known, estimated from a validation sample, or estimated from additional auxiliary data, the blip parameters will exhibit joint asymptotic normality. This asymptotic normality requires regularity conditions which are related to the previously discussed (Chapter 6) exceptional laws. As a reminder, exceptional laws were introduced by Robins [70], and are explored in depth in Appendix C.[8] In the presence of exceptional laws, several authors have considered techniques for reducing bias and correcting inference [56, 11, 15]. Theorem 7.7.1, presents results under the assumption of non-exceptional laws.

Theorem 7.7.1 (Asymptotic Normality of Modified G-Estimation)**.** *Suppose that for $j = 1, \ldots, K$ and $i = 1, \ldots, n$, we consistently estimate $P(A_{i,j}^*|H_{i,j}^*)$ and $\pi_j^*(H_{i,j}^*, A_{i,j}^*)$ through corresponding unbiased estimating equations, and we correctly specify the form of $C_j^*(H_{i,j}^*; \psi_j)$. Then the $(\widehat{\psi}_1, \ldots, \widehat{\psi}_K)$ which are estimated as components when solving $U_j^* = 0$ (Equation (7.7.1)) are asymptotically normal, under the independence assumptions from Theorem 7.5.1, and the regularity conditions set out by Robins [70] surrounding exceptional laws. Denoting $\widehat{\Psi} = (\widehat{\psi}_1, \ldots, \widehat{\psi}_K)$, we get that, as $n \to \infty$,*

$$\sqrt{n}\left(\widehat{\Psi} - \Psi\right) \xrightarrow{d} N\left(\mathbf{0}, \Sigma_\Psi\right).$$

Here $\Sigma_\Psi = I_\Psi \Sigma_\Theta I_\Psi$, I_Ψ is the diagonal matrix with 1's on the diagonal entries corresponding to the locations of the Ψ parameters in Θ, Θ is the solution to $E[U^*(\Theta)] = 0$, and Σ_Θ is sandwich variance matrix based on U^*.

[8] Briefly, exceptional laws correspond to those for which, with non-trivial probability, there is not a uniquely best treatment assignment.

The asymptotic normality, and as such asymptotic variance, follows directly from standard M-estimation theory, under the conditions outlined by Robins [70]. It is worth noting that, even under exceptional laws, G-estimation remains $\sqrt{n}$-consistent (as in Theorem 7.5.1), however, the limiting distribution depends on these laws. It is likely that the modified G-estimation procedure will benefit from the same types of bias corrections explored to account for exceptional laws. At present we ignore these considerations.

7.8 Prescribed, Actual, and Reported Treatments

The estimation procedure that has been outlined has assumed that we are making inference regarding the actual treatment (A) using the prescribed treatment (A^*). As introduced in Section 7.3 there are actually three treatments which may be available, though our discussion has not directly considered $A^\dagger$. As mentioned, an ITT based on A^* may have a valid causal interpretation but a naive analysis conducted with $A^\dagger$ in place of A does not.

The proposed correction functions equivalently whether a prescribed treatment, a reported treatment, or both are available for modelling. The challenge with using reported treatment in place of prescribed treatment is conceptual rather than mechanical. In order to model the misclassification probabilities, we require

$$\pi_j^\dagger(H_j^\dagger, A_j^\dagger) = P(A_j = 1 | A_j^\dagger, H_j^\dagger).$$

While there are no statistical concerns to specify and fit these models, this is a quantity which is harder to think about in terms of the actual subject matter. In these scenarios we are directly modelling the reverse causal direction. Similarly, the technique with prescribed treatments is aided by the ease with which $P(A_j^* = 1 | H_j^*)$ can be specified. Specifying a model for $P(A_j^\dagger = 1 | H_j^\dagger)$ is generally more challenging, owing to the less interpretable meaning of these quantities. If a model is specified for $P(A_j | H_j^\dagger)$, and a model is specified for $P(A_j^\dagger | A_j, H_j^\dagger)$, then these two quantities imply a model for $P(A_j^\dagger = 1 | H_j^\dagger)$. A similar technique could make simultaneous use of both $\{A_j^*, A_j^\dagger\}$.

If it is possible to specify the correct model for quantities based on reported treatments, and the relevant assumptions from Theorems 7.5.1 and 7.7.1 hold (with A_j^* replaced by $A_j^\dagger$, and similar substitutions), then the proposed correction procedure will exhibit the same consistency and asymptotic normality guarantees. This is a particularly powerful result as, to reiterate the point, a naive analysis replacing A_j with $A_j^\dagger$ has no causal interpretation whenever A_j has a true treatment effect, and A_j impacts $A_j^\dagger$.

7.9 Pseudo Outcomes and Optimal Treatments

In Theorem 7.5.1, the result depends on the construction of pseudo outcomes of the form

$$\widetilde{V}_{i,j} = \widetilde{V}_{i,j+1} + \left[A_{i,j}^{\text{opt}} - \pi_j^*(H_{i,j}^*)\right] C_j^*(H_{i,j}^*).$$

In Chapter 6 we discussed at length the difficult with constructing pseudo outcomes, even when the ψ_j parameters were consistently estimated. It is worth considering these pseudo outcomes in order to determine whether the same concerns exist. Under the assumptions of the theory that has been developed throughout this chapter, $\pi_j^*(H_{i,j}^*)$ and $C_j^*(H_{i,j}^*)$ are both known (up to parameters which are consistently estimated). As a result, consistent estimation relies entirely on whether or not $A_{i,j}^{\text{opt}}$ is consistently estimated, supposing that ψ_j is accurately estimated.

In the event that there are no treatment indicators in the blip function (which is to say, past treatments are not used as tailoring factors), then the blip function is exactly known when ψ_j is known, and $A_{i,j}^{\text{opt}}$ will be correctly specified. Otherwise, for any individual the optimal treatment regime may be incorrectly estimated based on the available data. This is because any of the A_ℓ contained as tailoring factors, with $\ell < j$ are replaced by $P(A_\ell = 1|H_j^*)$. Correspondingly, it is conceivable that the sign of C_j changes when moving from A_j to $E[A_j|\cdot]$.

Unlike in the case of errors in covariates assessing this possibility empirically can be done without much trouble. Suppose that we have ψ_j estimated correctly. As the only terms which are possibly error-prone in the blip function are the past treatment indicators, for any particular individual i, we can compute their optimal treatment assignment under different histories. Suppose, for sake of exposition, that for the j-th blip, only A_{j-1} is used as a tailoring factor. In Table 7.1 the possible combinations of optimal outcomes under the true treatment, as well as the predicted value using the misclassification model, are given. These results give a possible mechanism for empirically quantifying the degree to which suboptimal assignment may occur owing to nonadherence.

In terms of the impact that this possibly mischaracterized optimal treatment has on the consistency of the estimators we note that the key step of the proof is having

$$\nu_{k+1}^*(H_{k+1}^*) + \widetilde{A}_{k+1}^{\text{opt}} C_{k+1}^*(H_{k+1}^*) = E\left[\nu_{k+1}(H_{k+1}) + A_{k+1}^{\text{opt}} C_{k+1}(H_{k+1}) \middle| H_{k+1}^*, A_{k+1}^*\right].$$

Here we are using $\widetilde{A}_{k+1}^{\text{opt}}$ to represent the computed version of it, which was previously

Table 7.1: Scenarios corresponding to the possibility of mischaracterizing an individual's optimal response. The other two scenarios are not possible to realize. These probabilities can conceptually be computed for any particular individual, to determine for a given dataset the degree to which optimal treatments may be violated.

A_j^{opt} with $A_{j-1}=1$	A_j^{opt} with $A_{j-1}=0$	$\widehat{A}_j^{\text{opt}}$	$P(A_j^{\text{opt}} \neq \widehat{A}_j^{\text{opt}})$
1	1	1	0
1	0	1	$1-\pi_{j-1}^*(H_j^*)$
1	0	0	$\pi_{j-1}^*(H_j^*)$
0	1	1	$\pi_{j-1}^*(H_j^*)$
0	1	0	$1-\pi_{j-1}^*(H_j^*)$
0	0	0	0

assumed to exactly equal A_{k+1}^{opt}. This means that whenever we add a term which equals

$$E\left[A_{k+1}^{\text{opt}} C_{k+1}(H_{k+1}) \middle| H_{k+1}^*, A_{k+1}^*\right],$$

in place of $A_{k+1}^{\text{opt}} C_{k+1}^*(H_{k+1}^*)$ in the pseudo outcome, Equation (7.5.1), the resulting proof will hold. This expectation is computed under the assumption that ψ_{k+1} is known exactly, and that the form of $C_{k+1}(H_{k+1})$ is also precisely known. As before, whenever C_{k+1} does not depend on previous treatment indicators, this can be expressed as is. Otherwise, we can consider computing this conditional expectation based on the results of Table 7.1.

Suppose, again for the sake of expositional clarity,[9] that for C_{k+1}, only A_k is involved in the computation. Then, knowing the form of C_{k+1} we can say that

$$\begin{aligned}
&E\left[A_{k+1}^{\text{opt}} C_{k+1}(H_{k+1}) \middle| H_{k+1}^*, A_{k+1}^*\right] \\
&= P(A_k = 1 | H_{k+1}^*, A_{k+1}^*) E\left[A_{k+1}^{\text{opt}} C_{k+1}(H_{k+1}) \middle| H_{k+1}^*, A_{k+1}^*, A_k = 1\right] \\
&\quad + \left(1 - P(A_k = 1 | H_{k+1}^*, A_{k+1}^*)\right) E\left[A_{k+1}^{\text{opt}} C_{k+1}(H_{k+1}) \middle| H_{k+1}^*, A_{k+1}^*, A_k = 0\right] \\
&= \pi_k^*(H_{k+1}^*) I\left\{C_{k+1}(H_{k+1}^*, A_k = 1) > 0\right\} C_{k+1}(H_{k+1}^*, A_k = 1) \\
&\quad + \left[1 - \pi_k^*(H_{k+1}^*)\right] I\left\{C_{k+1}(H_{k+1}^*, A_k = 0) > 0\right\} C_{k+1}(H_{k+1}^*, A_k = 0).
\end{aligned}$$

In the event that the optimal treatment does not depend on A_k (either because A_k is not a term in the blip function, or because regardless of the value, the optimal treatment is the same), this will simplify to exactly $\widetilde{A}_{k+1}^{\text{opt}} C_{k+1}^*(H_{k+1}^*)$. Otherwise, this term is computable given the modelling that has been supposed. In this sense, it is possible to derive a pseudo

[9] If this is not the case, then the same argument holds conditioning on the complete past.

outcome which is valid regardless of whether or not previous treatment indicators are included. We can take

$$\widetilde{V}_j = \widetilde{V}_{j+1} + \pi^*_{j-1}(H^*_j) I\left\{C_j(H^*_j, A_j = 1) > 0\right\} C_j(H^*_j, A_j = 1) \\ + \left[1 - \pi^*_{j-1}(H^*_j)\right] I\left\{C_j(H^*_j, A_j = 0) > 0\right\} C_j(H^*_j, A_j = 0) - \pi^*_j(H^*_{i,j}) C^*_j(H^*_{i,j}).$$

While this results in a consistent estimate of a valid pseudo outcome under the outlined assumptions, it is worth noting that in practice the blip parameters will be estimated. Moreover, the misclassification probabilities are modelled, potentially based on external validation samples or previous studies. Correspondingly, the large sample properties gained through this added complexity may be undermined in some analyses based on the errors inherent to the estimation process. Fortunately, based on our previous discussions, it is possible to get a sense, empirically, of whether or not these pseudo outcomes are likely to materially differ from the previously suggested form in Equation 7.5.1. For the remainder of the Chapter, including for the simulation experiments, we will continue using the approximate pseudo outcomes, to demonstrate the utility of these expressions.

7.10 Multiple Treatment Alternatives

In this thesis thus far, G-estimation has been presented under the assumption that treatment is binary. The generalization of G-estimation to arbitrary categorical treatments is conceptually straightforward, complicating mostly the notation used. In place of assuming that

$$Q_j(H_j, A_j) = \nu_j(H_j) + A_j C_j(H_j; \psi_j),$$

the model instead is taken to be

$$Q_j(H_j, A_j) = \nu_j(H_j) + \sum_{A_{jk} \in \mathcal{A}_j} I(A_j = A_{jk}) C_{jk}(H_j; \psi_{jk}).$$

Here $\mathcal{A}_j$ is the set of possible treatment options at stage j, and C_{jk} is the k-th ($k = 1, \ldots, |\mathcal{A}_j|$) contrast function for stage j. We have

$$C_{jk}(H_j) = E[\widetilde{V}_{j+1} | H_j, A_j = a_{jk}] - E[\widetilde{V}_{j+1} | H_j, A_j = 0].$$

In the event of nonadherence, these quantities all naturally extend in the same was discussed before, replacing the contrast functions with their expected values, and replacing

the indicator function with the misclassification probabilities,

$$P(A_j = A_{jk}|A_j^*, H_j) = \pi_{jk}^*(H_j^*, A_j^*).$$

It is common for DTR estimation to be framed around binary treatments. However, in the event of nonadherence, it is important to question this framing. The presented theory above assumes that if an individual has been assigned treatment $A_j^* = 1$, and is nonadherent, then they are in the $A_j = 0$ category. Often, however, $A_j = 1$ will refer to a experimental treatment and $A_j = 0$ will refer to standard care. As a result, it may be the case that an individual who is nonadherent will not in fact be a member of the alternative treatment, but rather, an additional treatment category all together. If, in the scenario with complete adherence, we consider $A_j = 1$ to be an experimental treatment and $A_j = 0$ to be standard care, then it may be necessary to define a third category corresponding to no treatment. Then, if an individual does not adhere to $A_j^* = 1$ or $A_j^* = 0$, they are categorized in the third treatment category instead.

Consider the MACS analysis, for instance. In this case treatment refers to a decision to start Zidovudine (AZT) therapy. In this setting we will take $A_j = 1$ to refer to an individual starting AZT therapy at j, and $A_j = 0$ to mean that they have not taken AZT. Once prescribed the individual remains on AZT, and so the analyses we are basing our investigation on consider only timing of the therapy [96]. This setting is such that it is unlikely that there would be nonadherence when $A_j = 0$ is prescribed.[10] On the other hand, we know from the MACS data that adherence to $A_j = 1$ is not perfect. However, nonadherence to $A_j^* = 1$ is unlikely to result in $A_j = 0$ in all situations. Instead, partial adherence, wherein the treatment is taken but not according to all instructions, occurs more frequently. In this case we may wish to introduce a third category of actual treatment which corresponds to these partial compliers. Then we would observe either $A_j^* = 0$ or $A_j^* = 1$. If $A_j^* = 0$ then by assumption we take $A_j = 0$, however, if $A_j^* = 1$, we may have $A_j = 1$ (full adherence), $A_j = 0$ (full nonadherence), or $A_j = -1$ (partial adherence).

The precise categorizations for any setting will depend on the exact subject matter. In certain settings nonadherence will correspond exactly to the switching of treatments; in other situations, additional treatment options may need to be considered to more adequately represent the true, underlying reality. While doing this represents a more complex modelling setup, it is important to note that even if the true treatments were observable for all individuals, the added complexity would be necessary for valid causal conclusions to be drawn.

[10] AZT is a prescription drug, which is not likely to be readily accessed by those without a prescription.

7.11 Simulation Studies

7.11.1 Misclassification Dependent on Tailoring Variates

In the first experiment, we simulate a two-stage DTR with two primary tailoring variates, $X_1 \sim N(1,1)$ and $X_2 \sim N(1,4)$. At stage j treatment prescription depends only on X_j through a logistic model, with $P(A_j^* = 1|X_j) = \text{expit}(X_j)$. The true treatment depends on a similar logistic model, with $P(A_j = 1|A_j^*, X_j) \approx \text{expit}(-4.6 - 0.83X_j + 7.5A_j^*)$. The parameter values are taken to make the probability of misclassification fairly low, with values on average of 0.01 and 0.05 for those prescribed $A_j^* = 0$ and $A_j^* = 1$ respectively. The blip model at stage one is given by $1+X_1$, and at stage two it is given by $1+X_1+\psi_{22}A_1$, where $\psi_{22} \in \{-1, -0.1, 0, 0.1, 1\}$. The treatment free model is simply X_1, and the outcome has an error variance which is normally distributed with mean 0 and variance 2. We assume that there is a validation sample of 30% to estimate misclassification probabilities, and use a sample size of 1000. These simulations are repeated 1000 times. Box plots of the estimates across the replications comparing naive estimation (corresponding to an ITT), estimation using the true treatment (corresponding to an as-treated analysis), the proposed correction assuming misclassification probabilities are known, and the proposed correction assuming misclassification probabilities are estimated are shown in Figure 7.3.

The results indicate that the naive analysis produces clearly biased estimates of the parameters, while both corrections and the estimates based on the truth perform similarly. In some settings we see an improvement in the variability of the estimates for the corrected estimators compared to those which rely on the true treatment assignment.

7.11.2 Validation Set Sizing

In the second experiment, we maintain the same experimental setup as in the first set of analyses, however we fix $\psi_{22} = 0$. Instead, we vary the size of the replication set, based on a sample size of $n = 1000$, considering 10%, 20%, 30%, and 50% replication. These simulations were run with the same methods indicated above, though the estimates using the naive analysis, true analysis, and analysis based on known probabilities are independent of the size of the replication sample. As a result, in Figure 7.4, only box plots summarizing the results of the parameter estimates over the different sample sizes for the corrected estimator are shown.

While there is a trend towards decreasing variance, particularly moving from the 10% validation sample to the 20% validation sample, we see that the method performs similarly

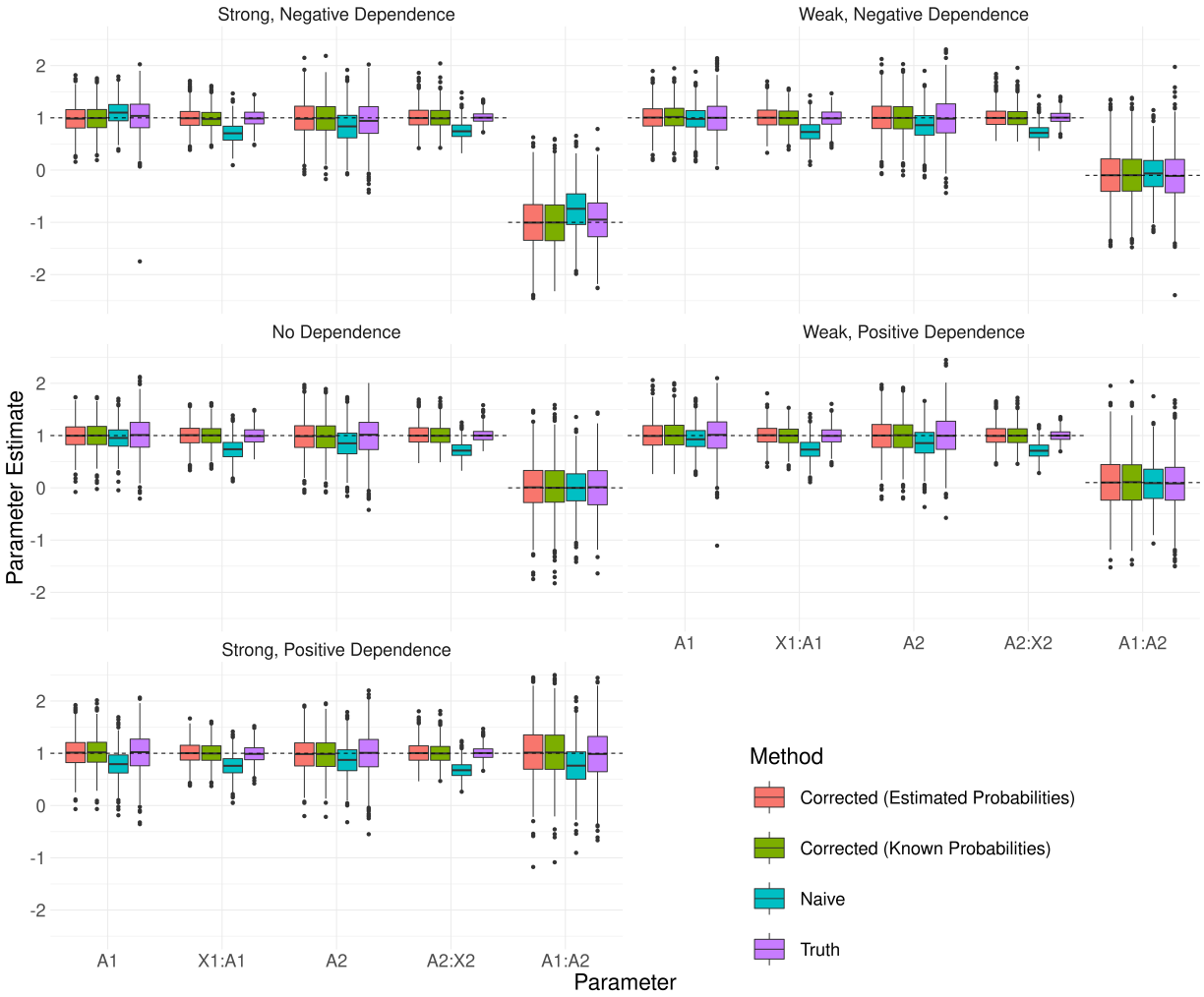

Figure 7.3: Estimated parameter values for the two-stage dynamic treatment regime varying the dependence of the second-stage on the tailoring effects of the misclassified A_1 variable. Each scenario compares the results of the correction (with known or estimated probabilities), the naive (ITT) analysis, and the as-treated analysis.

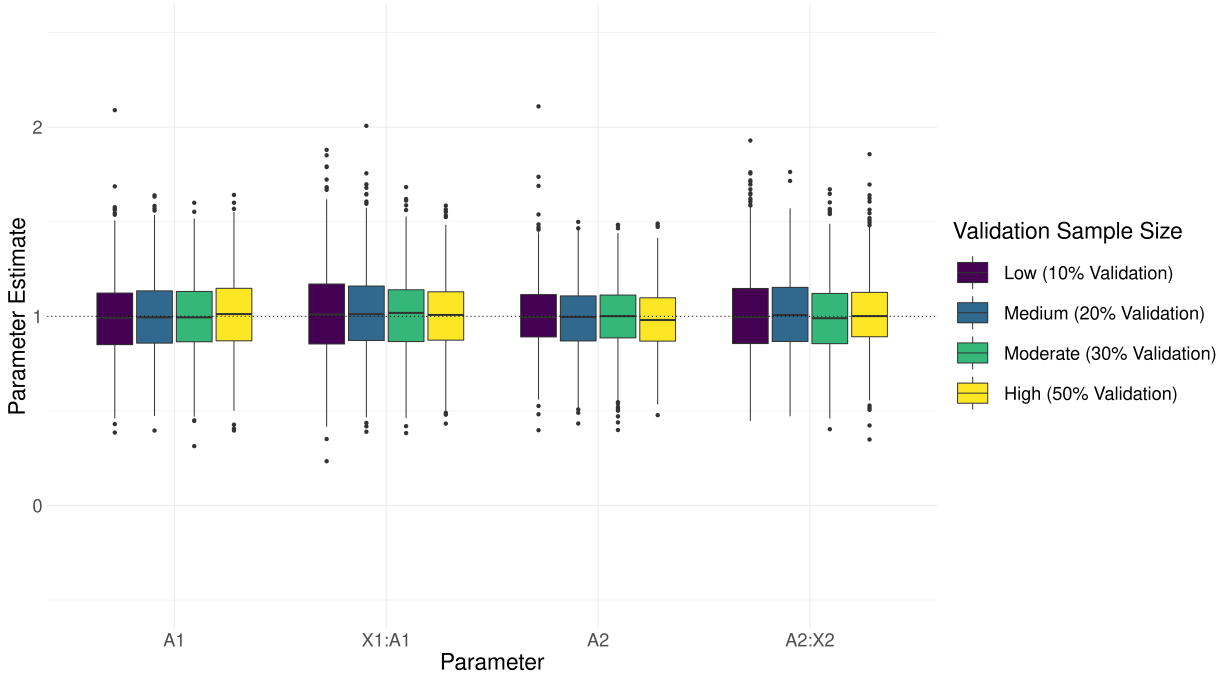

Figure 7.4: Estimated parameter values for the two-stage dynamic treatment regime varying the size of the validation sample. The dotted line indicates the true parameter values, with a sample size of $n = 1000$, and parameter estimates generated through the modified G-estimation procedure.

across all sizes of the validation sets considered.

7.11.3 Asymptotic Coverage Probabilities

In the third experiment, we consider the coverage probabilities obtained through the use of standard M-estimation theory. We simulate the same tailoring variables as in the previous scenarios. Treatment prescription is changed such that $P(A_1^* = 1|X_1) = \text{expit}(0.5 + X_1)$ while $P(A_2^* = 1|X_2) = \text{expit}(-0.5+X_2)$. The same misclassification probabilities are taken as in the first experiment. The jth blip model takes the form $1+X_j+\psi_{j2}A_j^*$, where $\psi_{12} = 1$ and $\psi_{22} = -1$. Further, the treatment free model becomes $X_1 + 0.5A_1^*$. We consider nine total scenarios where we vary the sample size to be low ($n = 200$), medium ($n = 1000$), or large ($n = 5000$) with the validation sample size being small (10%), medium (20%), or large (50%). We repeat each of these scenarios 1000 times, and consider the estimated standard errors for the blip terms that are based on approximate sandwich estimation techniques. Instead of explicitly solving the gradient of the estimating equation, we simply use numerical differentiation to approximate its value at the estimated parameter values. In Table 7.2 we include the number of simulations (out of 1000) which correctly covered the true parameter value using 90%, 95%, and 99% confidence intervals. In Figures 7.5 and 7.6, we plot the nominal significance threshold versus the estimated significance across all scenarios. These plots correspond to the empirical CDF (over the simulation replicates) of the p-values associated with testing whether the estimated parameter equals the true value, based on the estimated standard errors and a normal approximation. These plots are shown for the full range of α, in addition to $\alpha \in [0, 0.1]$.

The results suggest that coverage is well calibrated, even in small samples with a low validation percentage, though there is a notable improvement with increasing sample size as would be expected. The parameter ψ_{21} tended to exhibit the worst coverage behaviour, which is the blip parameter corresponding to X_2, however, coverage still tended to approximate the nominal levels well. It should be noted that the simulated scenario was not one under an exceptional law, and as such, standard asymptotic theory is expected to apply.

7.11.4 Reported Treatment Correction

In this simulation we consider the use of a reported treatment ($A^\dagger$) rather than a prescribed treatment. We once again look at a two-stage DTR. In this setting we have X_1, X_2 both

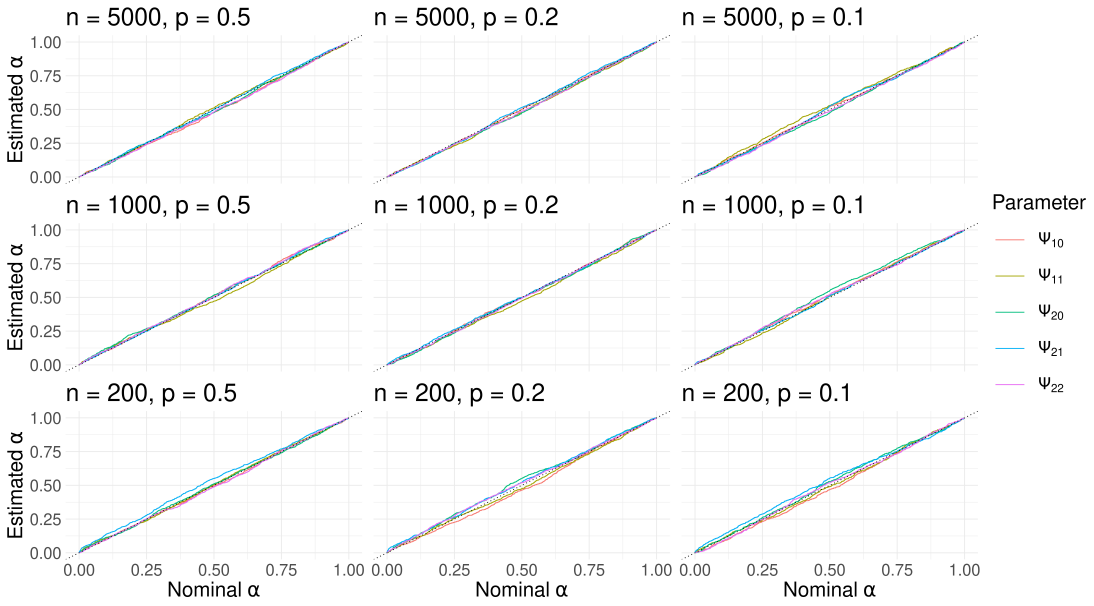

Figure 7.5: Plots of the empirical CDF of the p-values associated with testing whether the estimated blip parameter equals the true value, based on estimated standard errors and a normal approximation, for various sample sizes and validation proportions. The standard errors are formed based on numeric differentiation, within the modified G-estimation procedure.

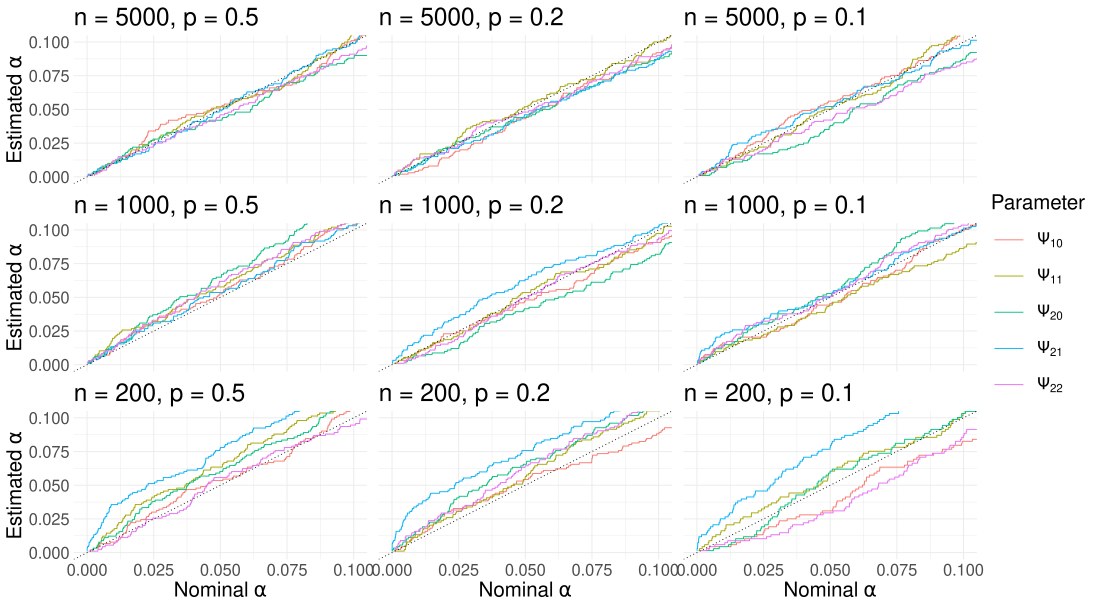

Figure 7.6: Plots of the empirical CDF of the p-values associated with testing whether the estimated blip parameter equals the true value, based on estimated standard errors and a normal approximation, for various sample sizes and validation proportions. This is the same results as in Figure 7.5, zoomed into $\alpha \in [0, 0.1]$. The standard errors are formed based on numeric differentiation, within the modified G-estimation procedure.

Table 7.2: Number of replications (out of 1000) which contained the true parameter value for in confidence intervals based on estimated standard errors, at the 90%, 95%, and 99% thresholds, for various sample sizes (n) and validation proportions (top rows 50%, middle rows 20%, bottom rows 10%). The confidence intervals are formed based on numeric differentiation, within the modified G-estimation procedure.

90% CI					95% CI					99% CI				
ψ_{10}	ψ_{11}	ψ_{20}	ψ_{21}	ψ_{22}	ψ_{10}	ψ_{11}	ψ_{20}	ψ_{21}	ψ_{22}	ψ_{10}	ψ_{11}	ψ_{20}	ψ_{21}	ψ_{22}
$n = 5000$														
899	894	912	901	909	948	948	958	951	954	990	989	990	990	988
908	900	913	912	910	957	948	956	954	952	995	987	990	991	986
894	886	913	903	916	944	950	962	948	958	987	992	993	987	989
$n = 1000$														
896	891	879	897	893	947	942	935	946	938	987	980	987	987	989
909	903	918	897	906	952	946	958	937	951	991	989	993	984	993
901	914	888	899	896	955	956	947	949	949	984	989	986	977	986
$n = 200$														
894	891	886	863	905	947	937	940	922	944	992	983	988	964	993
912	895	888	872	884	946	945	934	924	936	986	988	984	964	983
917	897	897	872	914	969	940	941	922	971	996	984	996	975	994

being discrete uniform on $\{-1, 0, 1\}$. Then, we define

$$P(A_j = 1|X_j) = \begin{cases} 0.2 & X = -1; \\ 0.5 & X = 0; \\ 0.8 & X = 1, \end{cases}$$

for $j = 1, 2$. Then we take

$$P(A_j^\dagger = 1|A_j, X_j) = \begin{cases} 0.95 & X_1 = -1; A_1 = 1; \\ 0.9 & X_1 = 0; A_1 = 1; \\ 0.85 & X_1 = 1; A_1 = 1; \\ 0.01 & X_1 = -1; A_1 = 0; \\ 0.05 & X_1 = 0; A_1 = 0; \\ 0.1 & X_1 = 1; A_1 = 0. \end{cases}$$

At stage 1 the blip is taken to be $1+X_1$, and at stage two it is taken as $1+\psi_{22}A_1$, where ψ_{22} is taken to vary across $\{-1, -0.1, 0, 0.1, 1\}$. The sample size is taken to be $n = 1000$, with 30% validation sample. In this setting we compare fitting the proposed correction, a naive (ITT) analysis, and an analysis which is based upon the true treatments. These simulations are repeated 1000 times. The MSE (multiplied by 100) are contained in Table 7.3.

Table 7.3: 100 times the observed MSE, based on 1000 replicated simulations, for the blip parameter estimates in a two-stage DTR, where reported treatments were used as a misclassified version of the truth. The results are based on a sample size of $n = 1000$, with a validation set of 30%, and they compare the corrected estimators, to those which naively apply G-estimation without correction, to those that are obtained when the true treatment status is reported.

	Corrected				Naive				Truth			
	ψ_{11}	ψ_{12}	ψ_{21}	ψ_{22}	ψ_{11}	ψ_{12}	ψ_{21}	ψ_{22}	ψ_{11}	ψ_{12}	ψ_{21}	ψ_{22}
$\psi_{22} = -1$	3.4	3.3	2.8	6.4	7.8	30.6	11.2	15.4	2.3	2.0	2.2	4.3
$\psi_{22} = -0.1$	3.6	2.8	2.7	5.5	12.5	22.5	7.2	4.2	2.0	1.9	2.1	3.9
$\psi_{22} = 0$	3.5	2.9	2.5	5.6	12.5	21.6	6.8	4.3	2.2	2.0	2.2	4.4
$\psi_{22} = 0.1$	3.4	3.2	2.6	5.9	13.0	21.2	6.7	4.8	2.1	2.1	2.3	4.8
$\psi_{22} = 1$	3.7	2.8	2.5	6.4	15.5	13.7	3.9	15.7	2.4	1.9	2.2	4.8

From these results we can see that, while the true estimators predictably have the lowest MSE across all of the presented scenarios, the corrected estimators perform comparably, despite the additional modelling requirements. The naive estimators exhibit large bias and greater variance, making them unreliable in general as a means of estimating the true treatment effectiveness. It is worth emphasizing that, unlike in the other scenarios where the ITT could (plausibly) be interpreted in a causal light, here not only are the naive estimators highly variable, they are also not defensible through any causal interpretation.

7.12 Multicenter AIDS Cohort Study (MACS) Analysis

Next we demonstrate the utility of our proposed corrections with an analysis of the Multicenter AIDS Cohort Study [45]. Our analysis primarily follows Wallace, Moodie, and Stephens [96] and Hernán, Brumback, and Robins [38]. MACS was a longitudinal study, which saw individuals twice a year, and at each visit survey questions and medical exams

were conducted. While the data are incredibly rich, our analysis focuses on a fairly simple question related to the treatment of HIV/AIDS. Our analysis seeks to estimate the optimal timing of intervention with a particular antiretroviral drug, Zidovudine (AZT), used to treat HIV/AIDS. AZT became available for the first time in March 1986, and so our analysis is restricted to only those individuals who were HIV-positive and AIDS-free, starting in March 1986. Because the primary purpose of this analysis is illustration of the application of these techniques, we further restrict our sample to include only the first two eligible visits, for any individual. Because patients were recruited in waves, the dates that the first two eligible visits take will differ across individuals.

The outcome of interest for the study will be the CD4 count for the individual at the visit following their second eligible visit. CD4 cells are white blood cells which are crucial for immune responses, and are commonly used to assess the health and progression of individual with HIV. In addition to using information regarding a patient's CD4 cell count, we will also take lab results regarding their CD8 counts, their white-blood cell (WBC) counts, their red blood cell (RBC) counts, their platelet counts, their blood pressure (systolic and diastolic), their weight, as well as a symptomatic indicator (which indicates whether the patients have had, at least one of the following symptoms, in their recent medical history: fever, oral candidiasis, diarrhea, weight loss, oral hairy leukoplakia, or herpes zoster). These variates were selected largely according to the analysis done by Hernán, Brumback, and Robins [38].

In October 1998 a questionnaire which assessed adherence to prescribed medication was added to the MACS [48]. This questionnaire assesses an individual's adherence to their prescribed regimen, over the previous four days. While it is generally the case that this self-reported adherence status may itself be misclassified, or not representative of typical behaviour from the patient in question, we ignore this in our analysis. That is, while this reported data are representative of $A^{\dagger}$, we treat is as though it were A. It is also worth acknowledging that, from the survey we can see that most patients who are not perfectly adherent to their prescribed treatment remain partially adherent. This fact is related to the discussion in Section 7.10 regarding the possibility of multiple treatment alternatives. A complete analysis of these data could make informed use of the levels of adherence that the patients report, classified based on the expected similarity of different doses. We will still make the common binary assumption.

Owing to the staggered entry into the study, the exclusion criteria previously discussed, and the use of the subset of publicly available data, the subset of individuals with adherence information forms a partially overlapping validation sample with the main survey. We will treat this data as though it is derived from an external validation sample, despite the fact that it is taken from the same study, and many individuals from the main study are

present in the validation set. Because of this the transportability requirements for using a validation sample are likely to be satisfied. One further note with regards to the adherence data is that, because this information only started to be collected in 1998, 12 years after eligibility into inclusion in our analysis, it is possible that adherence early in the study was subject to different forces than later when we have the information. We content ourselves with assuming that the adherence information is approximately representative of adherence throughout the study, but caution that this analysis is primarily useful insofar as it serves as an illustration of the proposed techniques.

By way of notation, we take $A_j^* \in \{0, 1\}$ to represent whether AZT was started at period j ($A_j^* = 1$) or not. We assume that once an individual has been prescribed AZT, they remain prescribed AZT, which means that once $A_j^* = 1$, there will be no change in this prescription. The true treatment, A_j, which is unobservable in general, corresponds to whether or not the individual took AZT during the j-th stage of the treatment. In terms of nonadherence we assume that if an individual has not been prescribed $A_j^* = 1$, then they will remain adherent. That is, $P(A_j = 0|A_j^* = 0) = 1$. The notation for each of the other predictors is summarized in Table 7.4.

Table 7.4: Set of predictor variables used in the analysis of the MACS data, along with their defined notation.

Variable	Description	
Birthdate	Year of Birth for the Individual	U
CD4 Count	Count of the number of CD4 cells present at visit j	C_j
CD8 Count	Count of the number of CD8 cells present at visit j	K_j
RBC Count	Count of the number of RBC cells present at visit j	R_j
WBC Count	Count of the number of WBC cells present at visit j	W_j
Platelet Count	Count of the number of platelets present at visit j	P_j
Systolic Blood Pressure	Systolic blood pressure measurement at stage j	T_j
Diastolic Blood Pressure	Diastolic blood pressure measurement at stage j	D_j
AIDS Status	Binary indicator of AIDS diagnosis at stage j	F_j
Body Weight	Individual's body weight in pounds at stage j	B_j
Symptom Indicator	Symptomatic status (as previously defined) at stage j	S_j

We begin by considering modelling the nonadherence directly on our (partially) external validation data. We specify a standard logistic regression model, and consider performing model selection through a combination of deviance tests and the BIC. The first question of interest in the modelling is whether or not the pattern of adherence appears to change over time. To do this we consider fitting the main effects model (with a logistic link) with

a log transformation of all factors except for U, S_j, and F_j, where we include a factor for each different visit present in the adherence dataset. The factors corresponding to the visit numbers are highly non-significant ($p \approx 0.99$ with a deviance test), and so we proceed assuming that a single nonadherence model can be used at each visit. From here, through a combination of BIC and deviance testing, we reduce the model until we are left with

$$\text{logit}(P(A_j = 1|\cdot)) = \alpha_0 + \alpha_1 U + \alpha_2 \log(D_j) + \alpha_3 \log(T_j) + \alpha_4 \log(C_j).$$

Note that this model has conditioned (implicitly) on $A_j^* = 1$, as we have assumed that $A_j^* = 0$ implies that $A_j = 0$. The resultant estimated coefficients are included in Table 7.5

Table 7.5: Parameter estimates (with standard errors, and Wald test statistics) from a logistic regression conducted on the external validation sample, estimating the propensity for those prescribed AZT to be (partially) nonadherent to their assigned treatment.

	Estimate	Standard Error	z value	Pr(>|z|)
(Intercept)	93.917	36.078	2.600	0.009
U	-0.049	0.018	-2.640	0.008
$\log(D_j)$	-4.073	1.494	-2.730	0.006
$\log(T_j)$	3.592	1.497	2.400	0.016
$\log(C_j)$	0.585	0.207	2.820	0.005

Taking this model for adherence, we can now begin fitting the dynamic treatment regime. Our analysis largely borrows from the findings of Wallace, Moodie, and Stephens [96] to inform the functional forms that are being considered. In our analysis, we consider only those individuals who have complete information (for the relevant factors), rather than conducting imputation. Further, owing to the size of the available data and the lack of subject-matter guidance, we consider the binary adherence mechanism previously outlined, rather than working through a model which adequately considers partial adherence. Despite these limitations the following analyses show the importance of addressing concerns with nonadherence, and can serve as a guide to a more thorough consideration of these topics.[11] In total, our models are fit using information from 2850 patients, representing information from a total of 8550 visits. The adherence information is based on 766 questionnaire responses, with a total of 220 patients providing this information. Of these 220 patients, 141 of them are included in the main sample; the remaining 79 do not have sufficient information from the required visits to be used in the main model fitting.

[11] That is to say, while the specific estimates from this analysis are subject to the aforementioned shortcomings, they do demonstrate the impact that nonadherence can have on optimal DTR estimation.

We take the previously discussed adherence model for both stages. For stage one we specify the outcome model as

$$\beta_{10} + \beta_{11}C_1 + \beta_{12}\log C_1 + A_1 \times (\psi_{11} + \psi_{12}U + \psi_{13}\log C_1 + \psi_{14}S_1)\,.$$

For the second stage, we take the outcome model to be given by

$$\beta_{20} + \beta_{21}C_1 + \beta_{22}\log C_1 + \beta_{23}C_2 + \beta_{24}\log C_2 + A_2 \times (\psi_{21} + \psi_{22}U + \psi_{23}\log C_2 + \psi_{24}S_2)\,.$$

For $j = 1, 2$ we specify the corresponding treatment prescription models as

$$\text{logit}(P(A_1^* = 1|\cdot)) = \gamma_{j1} + \gamma_{j2}C_j + \gamma_{j3}K_j + \gamma_{j4}R_j + \gamma_{j5}W_j + \gamma_{j6}P_j.$$

It is worth noting that, while larger blip models and treatment-free models were both considered, the variation in the available data made stable estimation of these treatment rules challenging. This was true whether conducting inference based on standard G-estimation (assuming that there was no nonadherence) or with the modified procedures. Instead, we use a simplified tailoring rule, more akin to that of Wallace, Moodie, and Stephens [96]. In an attempt to standardize the magnitude of coefficients, we transform U in these models to represent the patient's age in 1986, rather than their birth year.

With these models, we conduct both a naive analysis and one based on our proposed correction, with the specified nonadherence model. In order to assess the variability of these estimators, we conduct a bootstrap, based on 1000 replicates. The resulting point estimates and confidence intervals are displayed for both analyses in Table 7.6. From these estimates there are several points to notice. First, most effects are quite variable in both analyses: some of this would be remedied by standardizing the variables, which may allow for an easier interpretation. We can see that, while the magnitude of the corrected estimates tend to be larger, the point estimates seem to suggest the same directional effects across most of the factors. Of note are the results from ψ_{22} and ψ_{24}.

For both of these factors, corresponding to the tailoring effect of birth year and the presence of symptoms, respectively, the two approaches report differing impact at a 95% level of significance. Notably, the naive analysis would conclude that the impact of birth year does not differ substantially from 0, at a 95% level of significance, but that the presence of symptoms at stage two does. When correcting for adherence, these two points are reversed: birth year becomes a significant tailoring factor, while the presence of symptoms does not remain so. While it is interesting to note that there are differences in the tailoring factors that are estimated to have a significant impact on optimal treatment assignment, it is easier to see the influence that nonadherence has by considering optimal treatment

Table 7.6: Estimated blip parameters (with 95% bootstrapped confidence intervals) based on a naive analysis of MACS (assuming full adherence) compared with an analysis based on modified G-estimation procedure which accounts for the impacts of nonadherence.

	Naive			Corrected		
	Lower	Estimate	Upper	Lower	Estimate	Upper
ψ_{11}	-4647.81	8254.73	20884.43	-25434.32	40623.92	100192.05
ψ_{12}	-10.13	-3.93	2.50	-48.20	-19.51	12.76
ψ_{13}	-137.68	-49.61	29.90	-695.41	-182.41	171.44
ψ_{14}	-198.44	-82.76	46.39	-709.97	-193.30	284.67
ψ_{21}	-1065.59	5528.36	12594.28	3037.40	35922.73	66075.74
ψ_{22}	-6.14	-2.66	0.64	-32.62	-17.53	-1.34
ψ_{23}	-74.90	-34.41	8.85	-364.82	-120.10	174.50
ψ_{24}	-124.93	-70.34	-13.56	-545.72	-238.67	8.81

estimation. In Table 7.7 we present the results of estimating the optimal treatment across the entire dataset, based on each bootstrap iteration.

Table 7.7: The proportions of optimal treatment at stages one and two based on the 1000 bootstrap replicates for the MACS analysis. Presented here are the proportions (median across the replicates, as well as the minimum and maximum proportions) where the estimated optimal treatment agrees between the two analysis strategies, as well as the proportion of patients for whom treatment was recommended at each stage.

	Median	Minimum	Maximum
Stage One Optimal Treatment Agreement	0.913	0.083	1.000
Stage Two Optimal Treatment Agreement	0.962	0.356	1.000
Naive $\widehat{A}_1^{\text{opt}} = 1$	0.246	0.002	1.000
Naive $\widehat{A}_2^{\text{opt}} = 1$	0.049	0.000	0.804
Corrected $\widehat{A}_1^{\text{opt}} = 1$	0.203	0.010	1.000
Corrected $\widehat{A}_2^{\text{opt}} = 1$	0.085	0.000	0.828

In Table 7.7, we see that there tends to be a fairly high level of agreement between the two techniques (0.962 at the second stage, and 0.913 at the first stage), but this agreement is not perfect. It is worth pointing out that, within the adherence data we have access to, approximately 90% of respondents, who are assigned AZT treatment, are fully adherent. Moreover, in the data itself, roughly 2% and 5% of respondents were prescribed AZT at

each stage. Remember that we have assumed that all of those who were not prescribed were fully adherent. As a result, these 4% and 9% differences in optimal treatment assignment derive from an approximately half percentage of nonadherent patients in the data. The upper and lower bounds in the bootstrap replicates do make clear that these results are potentially highly variable, and it is worth reiterating that the specifics of this analysis may be subject to several shortcomings. Still, this analysis makes clear that even very small deviations from perfect adherence (0.5% in MACS) can have out-sized impacts on the ability to optimally treat patients. While it is unlikely that our specific effect estimates are perfectly indicative of the underlying reality, they do show the issue with ignoring adherence information in these contexts. It also demonstrates the caution required to ignore these types of impacts.

Chapter 8

Discussion

In this thesis, we explore concerns that arise, empirically, with common techniques used to correct for the effects of measurement error. We emphasize the important role that so-called *approximately consistent* error correction techniques play in reducing the bias of estimators when their grounding assumptions are met. These assumptions, while sometimes reasonable, are often violated in practice. Part I of this thesis explores how these assumptions can be relaxed without losing the attractive simplicity of the most commonly applied error correction techniques. Specifically, we concern ourselves with the situations where observed auxiliary data have error distributions that are evidently dissimilar, uniting some of the literature on replicate measurements with the literature on instrumental variable techniques. We also consider the impact of violations to normality on common measurement error correction techniques, and provide a nonparametric alternative which will be readily applied by analysts familiar with the existing techniques.

While the findings on these topics are presented as generally applicable to any analyses subject to measurement error, the initial motivation for relaxing these assumptions stemmed from observing data that are frequently used to estimate dynamic treatment regimes. Prior to this thesis, almost no work had been done to address the impacts of measurement error in DTR analyses – either within the tailoring covariates, or in the treatment indicators themselves. The estimation of an optimal DTR is a regression-based procedure and as such lends itself to some of the existing literature on errors in regression models. However, the available data made clear that generalizations to these existing techniques were necessary to have applicable methods on hand.

In addition to proposing generalizations to existing measurement error correction techniques, we also consider the problem of correcting for the impacts of measurement error

within dynamic treatment regimes. In Part II, we argue that naive analyses are generally not applicable, regardless of the framing of DTR estimation, and can lead to serious errors in inference. Moreover, we present techniques for overcoming these errors, which in following the theme of the thesis broadly, are designed to minimally impact an analyst's existing tools to apply. We address the issues that arise from errors in tailoring variables, adapting the work on the generalized methods presented in Part I, and we thoroughly explore issues that stem from nonadherence (or treatment misclassification). We demonstrate how these issues can be overcome using slight modifications to commonly applied correction techniques. Below, we summarize the key results contributed within this thesis.

Chapter 3

In Chapter 3 we focus predominantly on methods for relaxing the assumptions on the available auxiliary data that are used to perform measurement error corrections. We demonstrate how the commonly assumed structure of *replicate measurements*, where multiple proxies are assumed to be independent and identically distributed, will often be violated in practice. Frequently, it is the case that multiple proxy measurements are available, where each proxy may be subjected to a different error distribution. When this is the case, commonly applied techniques that rely on repeated measurements being identically distributed will range from inefficient to inconsistent. However, we also show that the assumption of identically distributed errors is unnecessary to continue to apply several common, approximately consistent error correction techniques.

We discuss how both regression calibration and simulation extrapolation can be expanded to this framework of repeated (rather than replicated) measurements, without otherwise changing the underlying process. The proposed estimators function in precisely the same way as the existing, commonly discussed techniques, whenever the repeated values are truly replicates. However, they also have the capacity to more efficiently make use of all of the observed data, and can lead to consistent corrections when these auxiliary data assumptions are violated.

In this chapter we prove consistent identification of the necessary moment parameters to perform a wide variety of corrections under this more general structure, and demonstrate that these estimators are asymptotically normal. We then show how these results can be combined by existing techniques – particularly regression calibration and SIMEX – to produce consistent and asymptotically normal corrections under a wider class of measurement error models than is typically assumed.

Chapter 4

Chapter 4 is primarily concerned with developing a deeper theoretical foundation for simulation extrapolation. Building on extensions to SIMEX, we explore the technique by framing it through the lens of functionals over the space of characteristic functions. In this way we are able to explore the asymptotic bias that arises from violations to the assumption of normally distributed errors. This framing presents a nonparametric generalization to the commonly discussed SIMEX, which we call the NP-SIMEX.

The NP-SIMEX exploits the empirical distribution to modify the simulation step of the standard simulation extrapolation. We show that, supposing the underlying estimator is sufficiently smooth, this procedure can be used regardless of the distribution of the errors, whenever validation data are available. In the presence of replicate measurements we need to assume that the errors follow a symmetric distribution, though, any symmetric distribution will do. We further demonstrate how, by invoking literature regarding kernel density estimation, the same procedure can be applied when errors are related to the underlying true measurement.

We prove that the NP-SIMEX results in consistent and asymptotically normal corrections, supposing some technical requirements on the estimators of interest. We demonstrate its applicability in a wide range of scenarios, and we discuss the principal drawback to the technique: namely, because it is nonparametric, it requires a substantial amount of data to be applied. When taken in conjunction with the theory developed around the analysis of the standard SIMEX, these results present a mechanism for assessing whether the errors appear sufficiently non-normal to warrant the use of the otherwise less efficient estimator.

Chapter 6

In Chapter 6, we consider dynamic treatment regimes, and begin to discuss the issues that measurement errors in tailoring variables present within this context. We break down the problem into the different roles that these tailoring variates play within the dynamic treatment regime and argue that through an application of regression calibration, to the dWOLS estimators, we can restore some of the desirable theoretical properties of the estimators.

This work builds on previously established work (in my Masters thesis) where the need for correction techniques was established. In this chapter, we further explore these arguments, and contend that even when the primary interest is in predicting optimal future treatment, it is worth considering corrections for the effects of errors.

We demonstrate that, whenever valid pseudo outcomes can be estimated, the proposed error correction techniques will result in doubly robust estimators for the true blip terms, under the considered measurement error models. This double robustness is a desirable property of dWOLS in the error-free setting. These estimators are shown to be asymptotically normal when we can make assumptions regarding the regularity of the law, as is commonly required within the DTR literature. The primary concern with the theoretical guarantees in this setting revolves around the need to generate valid pseudo outcomes. We discuss techniques that can be used, in certain settings, to develop such outcomes, and prove consistency of these estimators under restricted error models.

Chapter 7

In Chapter 7, we address the problem of nonadherence in dynamic treatment regimes. We argue that nonadherence is a problem that ought to be addressed in this framework, despite the frequent appeal to intention to treat analyses. We show how violations to the causal structure of a DTR can arise through nonadherence, in addition to the bias that is present when an ITT is used as a means of estimating the underlying truth. When combined with standard critiques of ITTs, this provides an argument for the need for intervention effectiveness methods for estimating optimal DTRs, even if they are to be used in conjunction with an ITT.

For the proposed correction we directly modify G-estimation, producing a doubly-robust, and asymptotically normal estimator, under standard regularity conditions. In this setting we explore different ways of modelling the nonadherence that may be present in data, and illustrate how different sources of auxiliary data can be used to facilitate the proposed correction. Moreover, we present the framework in such a way so as to be amenable to sensitivity analyses. The proposed estimators are demonstrated to work whether the proxy treatment indicator is an antecedent of the true treatment, or vice-versa, and we discuss how this framework can be made to apply when multiple treatment alternatives need to be considered.

The issue of pseudo outcomes, also discussed in Chapter 6, is further addressed within the context of nonadherence, where we demonstrate that in principle valid pseudo outcomes are estimable. This allows for, under fairly general assumptions, the consistent, doubly robust estimation of the true blip terms relating to the underlying treatment's efficacy.

Concluding Remarks and Future Work

In this thesis we present generalizations which successfully overcome several of the shortcomings in existing techniques to correct for the effects of measurement error, and building from here, propose the first substantive corrections for the effects of measurement error within the context of dynamic treatment regimes. We ground the methodologies in assumptions which are frequently observed in practice, justify their utility through theoretical arguments, and demonstrate their effectiveness through comprehensive simulation experiments. While these methods provide strong foundations for approaching the problems addressed within the thesis, there remain areas which are promising for future investigation.

The utility of the generalized error model presented in Chapter 3 extends beyond the demonstration of regression calibration and simulation extrapolation presented. While these techniques are broadly used, and as such useful to generalize, it seems likely to be the case that correction techniques which are catered to the specific setting of repeated rather than replicated measurements may prove more efficient. The specific estimators presented for regression calibration and simulation extrapolation are approximately consistent and are sensible estimators given the standard presentation, but no attempt at uncovering the optimal or most efficient estimators was made. Extending our discussion with a focus on estimator efficiency, or computational stability, may provide mechanisms for overcoming the primary shortcomings exhibited by the proposed methods. Our proposal of the nonparametric SIMEX in Chapter 4 theoretically relies on formal conditions that are difficult to check in practice; an attempt at re-characterizing the theory around conditions which are easier to verify would improve the utility of the techniques. As is common with nonparametric techniques, our results suggest that fairly large amounts of data are required to make the correction feasible. Theoretical results, or a wider empirical investigation, quantifying the exact impact of sample size on the efficiency of the technique would provide useful context for when the techniques may be applicable. As a final possible extension of the work on nonparametric simulation extrapolation, there appears to be a close relationship between the proposed techniques and the empirical simulation extrapolation. It may be worthwhile to investigate whether the similarities are more than aesthetic, and if so, whether the connections provide useful insight into the proposed technique.

Our investigation of errors in dynamic treatment regimes can be viewed as a preliminary investigation into these issues. There are several promising extensions both directly related to the proposed techniques, and moving beyond them. The results in Chapter 6 leveraging regression calibration for dWOLS estimators can likely be translated, with minimal modification, to estimators based on G-estimation; the same goes in reverse for the

nonadherence results in Chapter 7 being translated to dWOLS. The results regarding errors in variables are limited in two primary ways: first, only classical additive error was directly considered, and second, the construction of pseudo outcomes limits the capacity of the proposed estimators to achieve consistency. The first issue is likely able to be overcome by working directly with estimating equation approaches, which as demonstrated in Chapter 7, are particularly amenable to DTR estimation. The pseudo outcomes provide a barrier for regression-based techniques with a resolution which is less clear. It may be possible to leverage Q-learning, which uses a different formulation for the pseudo outcomes, to overcome these issues. More likely, however, classification-based techniques would provide a mechanism for overcoming these concerns more directly. It is worth noting, however, that a different set of trade-offs must be accepted to use non-regression based estimators. Chapters 6 and 7 also present the results as entirely separate; an investigation of techniques which can overcome both issues simultaneously, and perhaps investigate errors in the outcomes as well, would allow for more generally applicable techniques. The work presented in this thesis demonstrates the impact of errors on optimal DTR estimation, and just as a wide-ranging literature surrounding correction techniques for the effects of measurement error exist in other estimation settings, so too is there room for such a literature surrounding dynamic treatment regimes.

References

[1] Bailey, D. C. (2017). Not normal: the uncertainties of scientific measurements. *R. Soc. open sci.*, 4(1), 160600. doi:10.1098/rsos.160600.

[2] Barron, B. A. (1977). The effects of misclassification on the estimation of relative risk. *Biometrics*, 33(2), 414. doi:10.2307/2529795.

[3] Bhapkar, V. P. (1972). On a measure of efficiency of an estimating equation. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 34(4), 467–472.

[4] Bollinger, C. R. (1998). Measurement error in the current population survey: A nonparametric look. *Journal of Labor Economics*, 16(3), 576–594. doi:10.1086/209899.

[5] Braun, D., Gorfine, M., Parmigiani, G., Arvold, N. D., Dominici, F., and Zigler, C. (2017). Propensity scores with misclassified treatment assignment: a likelihood-based adjustment. *Biostatistics*, 18(4), 695–710.

[6] Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications (Chapman & Hall/CRC Interdisciplinary Statistics)*. Chapman and Hall/CRC.

[7] Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman and Hall/CRC.

[8] Carroll, R. J., Küchenhoff, H., Lombard, F., and Stefanski, L. A. (1996). Asymptotics for the simex estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, 91(433), 242–250.

[9] Carroll, R. J., Spiegelman, C. H., Lan, K. K. G., Bailey, K. T., and Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika*, 71(1), 19–25. doi:10.1093/biomet/71.1.19.

[10] Carroll, R. J. and Stefanski, L. A. (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85(411), 652–663. doi:10.1080/01621459.1990.10474925.

[11] Chakraborty, B., Laber, E. B., and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m -out-of- n bootstrap scheme. *Biometrics*, 69(3), 714–723. doi:10.1111/biom.12052.

[12] Chakraborty, B., Laber, E. B., and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3), 714–723.

[13] Chakraborty, B., Laber, E. B., and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptivem-out-of-nBootstrap scheme. *Biometrics*, 69(3), 714–723. doi:10.1111/biom.12052.

[14] Chakraborty, B. and Moodie, E. E. (2013). *Statistical Methods for Dynamic Treatment Regimes*. Springer New York. doi:10.1007/978-1-4614-7428-9.

[15] Chakraborty, B., Murphy, S., and Strecher, V. (2009). Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat Methods Med Res*, 19(3), 317–343. doi:10.1177/0962280209105013.

[16] Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428), 1314–1328. doi:10.1080/01621459.1994.10476871.

[17] Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Federation Proceedings*, 21(4)Pt 2, 58–61.

[18] Cotton, C. A. and Heagerty, P. J. (2011). A data augmentation method for estimating the causal effect of adherence to treatment regimens targeting control of an intermediate measure. *Stat Biosci*, 3(1), 28–44. doi:10.1007/s12561-011-9038-1.

[19] Council, N. R. (1986). *Nutrient Adequacy*. National Academies Press. doi:10.17226/618.

[20] de Angelis, D. and Young, G. A. (1992). Smoothing the bootstrap. *International Statistical Review / Revue Internationale de Statistique*, 60(1), 45. doi:10.2307/1403500.

[21] Devanarayan, V. and Stefanski, L. A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics & Probability Letters*, 59(3), 219–225. doi:10.1016/s0167-7152(02)00098-6.

[22] DiMatteo, M. R. (2004). Variations in patients' adherence to medical recommendations. *Medical Care*, 42(3), 200–209. doi:10.1097/01.mlr.0000114908.90348.f9.

[23] Durbin, J. (1960). Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 22(1), 139–153.

[24] Eckert, R. S., Carroll, R. J., and Wang, N. (1997). Transformations to additivity in measurement error models. *Biometrics*, 53(1), 262–272.

[25] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611970319.

[26] Fava, M., Rush, A., Trivedi, M. H., Nierenberg, A. A., Thase, M. E., Sackeim, H. A., Quitkin, F. M., Wisniewski, S., Lavori, P. W., Rosenbaum, J. F., Kupfer, D. J., and for the STAR∗D Investigators Group (2003). Background and rationale for the sequenced treatment alternatives to relieve depression (star∗d) study. *Psychiatric Clinics of North America*, 26(2), 457–494. doi:10.1016/s0193-953x(02)00107-7.

[27] Fernholz, L. T. (1983). *von Mises Calculus For Statistical Functionals*. Springer New York. doi:10.1007/978-1-4612-5604-5.

[28] Filippova, A. A. (1962). Mises' theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. *Theory of Probability & Its Applications*, 7(1), 24–57. doi:10.1137/1107003.

[29] Freedman, L. S., Fainberg, V., Kipnis, V., Midthune, D., and Carroll, R. J. (2004). A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics*, 60(1), 172–181. doi:10.1111/j.0006-341X.2004.00164.x.

[30] Fuller, W. A., editor (1987). *Measurement Error Models*. John Wiley & Sons, Inc. doi:10.1002/9780470316665.

[31] Gleser, L. J. (1990). Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. In *Statistical analysis of measurement error models and applications (Arcata, CA, 1989)*, volume 112 of *Contemp. Math.*, 99–114. Amer. Math. Soc., Providence, RI. doi:10.1090/conm/112/1087101.

[32] Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.*, 31(4), 1208–1211. doi:10.1214/aoms/1177705693.

[33] Gonzalez, J. S., Batchelder, A. W., Psaros, C., and Safren, S. A. (2011). Depression and HIV/AIDS treatment nonadherence: A review and meta-analysis. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 58(2), 181–187. doi: 10.1097/qai.0b013e31822d490a.

[34] Gupta, S. (2011). Intention-to-treat concept: A review. *Perspect Clin Res*, 2(3), 109. doi:10.4103/2229-3485.83221.

[35] Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468), 1015–1026. doi:10.1198/016214504000000548.

[36] Han, S. (2021). Identification in nonparametric models for dynamic treatment effects. *Journal of Econometrics*, 225(2), 132–147. doi:10.1016/j.jeconom.2019.08.014.

[37] Hansen, R. A., Kim, M. M., Song, L., Tu, W., Wu, J., and Murray, M. D. (2009). Adherence: Comparison of methods to assess medication adherence and classify nonadherence. *Ann Pharmacother*, 43(3), 413–422. doi:10.1345/aph.1l496.

[38] Hernán, M. Á., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology*, 11(5), 561–570. doi:10.1097/00001648-200009000-00012.

[39] Hernan, M. A., Lanoy, E., Costagliola, D., and Robins, J. M. (2006). Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology & Toxicology*, 98(3), 237–242. doi:10.1111/j.1742-7843.2006.pto_329.x.

[40] Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 221–233. University of California Press, Berkeley, Calif.

[41] Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. John Wiley & Sons, Inc. doi:10.1002/9780470434697.

[42] Kallianpur, G. (1963). Von mises functionals and maximum likelihood estimation. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 25(2), 149–158.

[43] Kallianpur, G. and Rao, C. R. (1955). On fisher's lower bound to asymptotic variance of a consistent estimate. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 15(4), 331–342.

[44] Kannel, W., Neaton, J., Wentworth, D., Thomas, H., Stamler, J., Hulley, S., and Kjelsberg, M. (1986). Overall and coronary heart disease mortality rates in relation to major risk factors in 325, 348 men screened for the MRFIT. *American Heart Journal*, 112(4), 825–836. doi:10.1016/0002-8703(86)90481-3.

[45] Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and and, C. R. R. (1987). The multcenter AIDS cohort study: Rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology*, 126(2), 310–318. doi:10.1093/aje/126.2.310.

[46] Khudyakov, P., Gorfine, M., Zucker, D., and Spiegelman, D. (2015). The impact of covariate measurement error on risk prediction. *Statist. Med.*, 34(15), 2353–2367. doi:10.1002/sim.6498.

[47] Kipnis, V. (2003). Structure of dietary measurement error: Results of the OPEN biomarker study. *American Journal of Epidemiology*, 158(1), 14–21. doi:10.1093/aje/kwg091.

[48] Kleeberger, C. A., Phair, J. P., Strathdee, S. A., Detels, R., Kingsley, L., and Jacobson, L. P. (2001). Determinants of heterogeneous adherence to HIV-antiretroviral therapies in the multicenter AIDS cohort study. *JAIDS Journal of Acquired Immune Deficiency Syndromes Journal of Acquired Immune Deficiency Syndromes*, 26(1), 82–92. doi:10.1097/00126334-200101010-00012.

[49] Kosorok, M. R. and Laber, E. B. (2019). Precision medicine. *Annu. Rev. Stat. Appl.*, 6(1), 263–286. doi:10.1146/annurev-statistics-030718-105251.

[50] Koul, H. L. and Song, W. (2014). Simulation extrapolation estimation in parametric models with laplace measurement error. *Electron. J. Statist.*, 8(2), 1973–1995. doi:10.1214/14-EJS941.

[51] Küchenhoff, H., Mwalili, S. M., and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification simex. *Biometrics*, 62(1), 85–96.

[52] Li, F., Morgan, K. L., and Zaslavsky, A. M. (2017). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521), 390–400. doi:10.1080/01621459.2016.1260466.

[53] Marshall, R. J. (1990). Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *Journal of Clinical Epidemiology*, 43(9), 941–947. doi:10.1016/0895-4356(90)90077-3.

[54] McCoy, E. (2017). Understanding the intention-to-treat principle in randomized controlled trials. *WestJEM*, 18(6), 1075–1078. doi:10.5811/westjem.2017.8.35985.

[55] McKenzie, H. W., Jerde, C. L., Visscher, D. R., Merrill, E. H., and Lewis, M. A. (2008). Inferring linear feature use in the presence of GPS measurement error. *Environ Ecol Stat*, 16(4), 531–546. doi:10.1007/s10651-008-0095-7.

[56] Moodie, E. E. M. and Richardson, T. S. (2010). Estimating optimal dynamic regimes: Correcting bias under the null. *Scandinavian Journal of Statistics*, 37(1), 126–146. doi:10.1111/j.1467-9469.2009.00661.x.

[57] Morrissey, M. J. and Spiegelman, D. (1999). Matrix methods for estimating odds ratios with misclassified exposure data: Extensions and comparisons. *Biometrics*, 55(2), 338–344. doi:10.1111/j.0006-341x.1999.00338.x.

[58] Morton, R. (1981). Efficiency of estimating equations and the use of pivots. *Biometrika*, 68(1), 227–233. doi:10.1093/biomet/68.1.227.

[59] Murphy, S. A. (2003). Optimal dynamic treatment regimes. *J Royal Statistical Society, Series B*, 65(2), 331–355. doi:10.1111/1467-9868.00389.

[60] Murray, R. P., Connett, J. E., Lauger, G. G., and Voelker, H. T. (1993). Error in smoking measures: effects of intervention on relations of cotinine and carbon monoxide to self-reported smoking. the lung health study research group. *Am J Public Health*, 83(9), 1251–1257. doi:10.2105/ajph.83.9.1251.

[61] Nakamura, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika*, 77(1), 127–137. doi:10.1093/biomet/77.1.127.

[62] National Heart, Lung, and Blood Institute (2019). Framingham heart study teaching dataset. https://biolincc.nhlbi.nih.gov/teaching/. Accessed: 2021-09-16.

[63] Novick, S. J. and Stefanski, L. A. (2002). Corrected score estimation via complex variable simulation extrapolation. *Journal of the American Statistical Association*, 97(458), 472–481.

[64] Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association*, 91(436), 1440–1449. doi: 10.1080/01621459.1996.10476712.

[65] Purdom, E. and Holmes, S. P. (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 4(1). doi: 10.2202/1544-6115.1070.

[66] Rajan, A. and Desai, S. (2018). Non-gaussian error distributions of galactic rotation speed measurements. *Eur. Phys. J. Plus*, 133(3). doi:10.1140/epjp/i2018-11946-7.

[67] Ranganathan, P., Pramesh, C., and Aggarwal, R. (2016). Common pitfalls in statistical analysis: Intention-to-treat versus per-protocol analysis. *Perspect Clin Res*, 7(3), 144. doi:10.4103/2229-3485.184823.

[68] Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9-12), 1393–1512. doi:10.1016/0270-0255(86)90088-6.

[69] Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Lecture Notes in Statistics*, 69–117. Springer New York. doi:10.1007/978-1-4612-1842-5_4.

[70] Robins, J. M. (2004). *Optimal Structural Nested Models for Optimal Sequential Decisions*, 189–326. Springer New York, New York, NY. doi:10.1007/978-1-4419-9076-1_11.

[71] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701. doi: 10.1037/h0037350.

[72] Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.*, 6(1), 34–58. doi:10.1214/aos/1176344064.

[73] Rubin, D. B. (1980). Bias reduction using mahalanobis-metric matching. *Biometrics*, 36(2), 293. doi:10.2307/2529981.

[74] Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469), 322–331. doi:10.1198/016214504000001880.

[75] Rudemo, M., Ruppert, D., and Streibig, J. C. (1989). Random-effect models in nonlinear regression with applications to bioassay. *Biometrics*, 45(2), 349–362.

[76] Rush, A., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., Thase, M. E., Nierenberg, A. A., Quitkin, F. M., Kashner, T., Kupfer, D. J., Rosenbaum, J. F., Alpert, J., Stewart, J. W., McGrath, P. J., Biggs, M. M., Shores-Wilson, K., Lebowitz, B. D., Ritz, L., Niederehe, G., and for the STAR∗D Investigators Group (2004). Sequenced treatment alternatives to relieve depression (star∗d): rationale and design. *Controlled Clinical Trials*, 25(1), 119–142. doi: 10.1016/s0197-2456(03)00112-0.

[77] Schaalje, G. B. and Butts, R. A. (1993). Some effects of ignoring correlated measurement errors in straight line regression and prediction. *Biometrics*, 49(4), 1262. doi:10.2307/2532270.

[78] Serfling, R. (1980). *Approximation theorems of mathematical statistics*. Wiley, New York.

[79] Shalizi, C. R. (2021). Advanced data analysis from an elementary point of view. Unpublished. Accessed as of May 2022.

[80] Shaw, P. A., Deffner, V., Keogh, R. H., Tooze, J. A., Dodd, K. W., Küchenhoff, H., Kipnis, V., and Freedman, L. S. (2018). Epidemiologic analyses with error-prone exposures: review of current practice and recommendations. *Annals of Epidemiology*, 28(11), 821–828. doi:10.1016/j.annepidem.2018.09.001.

[81] Sheiner, L. B. and Rubin, D. B. (1995). Intention-to-treat analysis and the goals of clinical trials. *Clin Pharmacol Ther*, 57(1), 6–15. doi:10.1016/0009-9236(95)90260-0.

[82] Simoneau, G., Moodie, E. E. M., Platt, R. W., and Chakraborty, B. (2017). Non-regular inference for dynamic weighted ordinary least squares: understanding the impact of solid food intake in infancy on childhood weight. *Biostatistics*, 19(2), 233–246. doi:10.1093/biostatistics/kxx035.

[83] Smith, M. K., Bracken, D. S., Rudy, C. R., and Santi, P. A. (Jan 2005). An analysis of the systematic components of calorimetry uncertainty. Technical report, Los Alamos National Laboratory, United States. LA-UR–05-4397.

[84] Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4), 465–472.

[85] Stefanski, L. A. and Boos, D. D. (2002). The calculus of m-estimation. *The American Statistician*, 56(1), 29–38.

[86] Stefanski, L. A. and Buzas, J. S. (1995). Instrumental variable estimation in binary regression measurement error models. *Journal of the American Statistical Association*, 90(430), 541–550.

[87] Stefanski, L. A. and Cook, J. R. (1995). Simulation-extrapolation: The measurement error jackknife. *Journal of the American Statistical Association*, 90(432), 1247–1256.

[88] The Japanese Society of Hypertension (2009). Measurement and clinical evaluation of blood pressure. *Hypertension Research*, 32(1), 11–23. doi:10.1038/hr.2008.2.

[89] Thiébaut, A. C., Freedman, L. S., Carroll, R. J., and Kipnis, V. (2007). Is it necessary to correct for measurement error in nutritional epidemiology? *Ann Intern Med*, 146(1), 65. doi:10.7326/0003-4819-146-1-200701020-00012.

[90] Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. (2019). *Dynamic Treatment Regimes*. Chapman and Hall/CRC. doi:10.1201/9780429192692.

[91] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press. doi:10.1017/cbo9780511802256.

[92] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer New York. doi:10.1007/978-1-4757-2545-2.

[93] Villanueva, E. V. (2001). The validity of self-reported weight in US adults: a population based cross-sectional study. *BMC Public Health Public Health*, 1(1). doi: 10.1186/1471-2458-1-11.

[94] Wallace, M. P. (2022). Measurement error and precision medicine. In T. Cai, B. Chakraborty, E. Laber, E. Moodie, and M. van der Laan, editors, *Handbook of Statistical Methods for Precision Medicine [Accepted]*, Handbooks of Modern Statistical Methods. Chapman and Hall/CRC, Boca Raton Florida.

[95] Wallace, M. P. and Moodie, E. E. M. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics*, 71(3), 636–644. doi: 10.1111/biom.12306.

[96] Wallace, M. P., Moodie, E. E. M., and Stephens, D. A. (2016). Model assessment in dynamic treatment regimen estimation via double robustness. *Biom*, 72(3), 855–864. doi:10.1111/biom.12468.

[97] Walter, S. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology*, 41(9), 923–937. doi:10.1016/0895-4356(88)90110-2.

[98] Watkins, C. J. C. H. and Dayan, P. (1992). Q-learning. *Mach Learn*, 8(3-4), 279–292. doi:10.1007/bf00992698.

[99] Xu, Y., Kim, J. K., and Li, Y. (2017). Semiparametric estimation for measurement error models with validation data. *Canadian Journal of Statistics*, 45(2), 185–201.

[100] Xu, Y., Kim, J. K., and Li, Y. (2017). Semiparametric estimation for measurement error models with validation data. *Can. J. Statistics*, 45(2), 185–201. doi:10.1002/cjs.11314.

[101] Yi, G. (2017). *Statistical Analysis With Measurement Error or Misclassification: Strategy, Method and Application*. Springer, New York.

[102] Yi, G. Y. and He, W. (2012). Bias analysis and the simulation-extrapolation method for survival data with covariate measurement error under parametric proportional odds models. *Biometrical Journal*, 54(3), 343–360. doi:10.1002/bimj.201100037.

[103] Zolnierek, K. B. H. and DiMatteo, M. R. (2009). Physician communication and patient adherence to treatment. *Medical Care*, 47(8), 826–834. doi:10.1097/mlr.0b013e31819a5acc.

APPENDICES

Appendix A

M-Estimation Supplement

In this appendix chapter we provide an overview of the necessary results of M-estimators that we make use of.

A.1 Background and Setup

When we have a general $p \times 1$ parameter, Θ, which is of interest related to the distribution of a random variable Z, we are typically concerned with finding an estimator $\widehat{\Theta}$ that is a function of an iid random sample, say $Z_1, \ldots, Z_n$. Many such estimators can be expressed as the solution to a set of estimating equations, represented as

$$U_n(\widehat{\Theta}) = \sum_{i=1}^{n} \Psi(Z_i; \widehat{\Theta}) = 0.$$

Here, the notation $U_n(\widehat{\Theta})$ emphasizes the fact that ultimately this is a function of an estimated parameter value, after we have fixed the random sample. Such estimators are called *M-estimators*.[1] The function $\Psi(Z; \Theta)$ is called an *unbiased estimating equation* if $E_\Theta [\Psi(Z; \Theta)] = 0$.

Formally, if $p(Z; \Theta)$ is the density of Z with respect to ν, when Θ is the relevant parameter, then $E_\Theta[\Psi(Z; \Theta)] = \int \Psi(z; \Theta) p(Z; \Theta) d\nu(z)$. We will typically suppress this notation where it can be inferred without confusion. If Ψ is an unbiased estimating equation, then

[1] More broadly, M-estimators are the zeros of estimating functions, but we use this language interchangably.

$\widehat{\Theta}$ will be endowed with certain desirable properties under sufficient regularity conditions. Namely, it will be consistent and asymptotically normal (CAN). If Θ_0 is the true value for the parameter, then $\widehat{\Theta} \xrightarrow{p} \Theta_0$, and $\sqrt{n}\left(\widehat{\Theta} - \Theta_0\right) \xrightarrow{d} N(0, \Sigma)$.

The asymptotic covariance Σ is given by

$$\Sigma = \left\{ E\left[\frac{\partial}{\partial \Theta'} \Psi(Z;\Theta)\Big|_{\Theta=\Theta_0} \right] \right\}^{-1} E\left[\Psi(Z;\Theta_0)\Psi(Z;\Theta_0)'\right] \left\{ E\left[\frac{\partial}{\partial \Theta'} \Psi(Z;\Theta)\Big|_{\Theta=\Theta_0} \right] \right\}^{-1'}.$$

This is often denoted by $\Sigma = \mathcal{A}(\Theta_0)^{-1}\,\mathcal{B}(\Theta_0)\,\mathcal{A}(\Theta_0)^{-1'}$, and referred to as the *sandwich formula*. For any specifically chosen Ψ, it is possible to compute both

$$\widehat{\mathcal{A}}(\Theta^*) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \Theta'} \Psi(Z_i;\Theta)\Big|_{\Theta=\Theta^*}, \text{ and}$$
$$\widehat{\mathcal{B}}(\Theta^*) = \frac{1}{n}\sum_{i=1}^{n} \Psi(Z_i;\Theta^*)\Psi(Z_i;\Theta^*)'.$$

Using the law of large numbers it is clear that $\widehat{\mathcal{A}}(\Theta^*) \xrightarrow{p} \mathcal{A}(\Theta^*)$, and that $\widehat{\mathcal{B}}(\Theta^*) \xrightarrow{p} \mathcal{B}(\Theta^*)$. This gives the motivation for the *sandwich estimator* for the variance, given by $\widehat{\Sigma} = \widehat{\mathcal{A}}(\widehat{\Theta})^{-1}\widehat{\mathcal{B}}(\widehat{\Theta})\widehat{\mathcal{A}}(\widehat{\Theta})^{-1'}$. Combining this with the asymptotic distribution, we get that the approximate sampling distribution of $\widehat{\Theta}$ is $N(\Theta_0, n^{-1}\widehat{\Sigma})$, allowing for standard confidence intervals to be computed.

This general theory provides the framework for the estimation procedures. There are many sets of regularity conditions which will suffice for this theory to hold. In the following, we illustrate what is generally required of Ψ in order for this to be the case.

A.2 Regularity Conditions

If $E[\Psi(Z;\Theta_0)] = 0$ does not uniquely determine Θ_0, then in general there will be concerns with the technique. If it does, however, then there exists a sequence of M-estimators, $\widehat{\Theta}$, which are consistent for Θ_0 [40, 78]. The remainder of the asymptotic derivations rely on a Taylor series expansion of $U_n(\Theta)$. As a result, we need Ψ to be sufficiently smooth to allow for the Taylor expansion. Taking $\dot{U}_n(\Theta^*) = \frac{\partial}{\partial \Theta'} U_n(\Theta)\big|_{\Theta=\Theta^*}$, then since $\widehat{\Theta}$ is a root

of $U_n(\cdot)$, we write

$$0 = U_n(\widehat{\Theta}) = U_n(\Theta_0) + \dot{U}_n(\Theta_0)(\widehat{\Theta} - \Theta_0) + R_n.$$

Assuming that $\dot{U}_n(\Theta_0)$ is non-singular, we can re-arrange to have

$$\sqrt{n}\left(\widehat{\Theta} - \Theta_0\right) = \sqrt{n}\left[-\dot{U}_n(\Theta_0)\right]^{-1} U_n(\Theta_0) + \sqrt{n}R_n^*,$$

where $R_n^* = \left[-\dot{U}_n(\Theta_0)\right]^{-1} R_n$.

If $\mathcal{A}(\Theta_0)$ exists, then the weak law of large numbers will give $-\dot{U}_n(\Theta_0) \xrightarrow{p} -\mathcal{A}(\Theta_0)$. If $\mathcal{B}(\Theta_0)$ exists, then $\sqrt{n}U_n(\Theta_0) \xrightarrow{d} N(0, \mathcal{B}(\Theta_0))$, by the central limit theorem. Now, so long as $R_n^* = o_p(n^{-1/2})$, then $\sqrt{n}R_n^* \xrightarrow{p} 0$. If this does in fact hold then a straightforward application of Slutsky's Theorem gives the necessary distributional results. In both Huber [40] and Serfling [78], conditions for this are given, however, it will generally be the case that if Ψ is smooth and Θ does not grow in dimension quickly as $n \to \infty$, it will be the case. Summarizing, we have the following conditions

1. Θ_0 is uniquely identified by the unbiased estimating equation. This gives consistency.
2. $\dot{U}_n(\Theta_0)$ is non-singular, which should happen for sufficiently large n.
3. $\mathcal{A}(\Theta_0)$ and $\mathcal{B}(\Theta_0)$ exist, so that the WLLN and CLT can be applied.
4. $R_n^* = o_p(n^{-1/2})$. This will roughly hold when Ψ is smooth and Θ is a fixed dimension.

These regularity conditions will allow for the asymptotic theory presented previously to hold.

While the regularity conditions are fairly mild, there are some important cases where they will be violated. Of particular interest for this thesis, if Ψ contains an indicator function which depends on Θ, this will make Ψ sufficiently non-smooth so as to render standard M-estimator theory invalid. This is fundamentally the issue with applying M-estimator theory to DTR estimation directly, since the pseudo-outcomes are typically calculated using a maximization function. This is not a problem if the indicator functions contained in Ψ do not contain Θ, as would be the case if, for instance, we use indicators regarding whether or not X_{ij}^* is observed.

Appendix B

Theoretical Results

B.1 Chapter 3

Proof of Lemma 3.5.1. We will prove this result under the assumption of incomplete replication and with the measured covariate Z. First, we note that for any observed variable A, we can get consistent estimates of the mean and variance of A (denoted μ_A and Σ_A) by solving the estimating equation

$$0 = \sum_{i=1}^{n} \begin{pmatrix} A_i - \mu_A \\ (A_i - \mu_A)^2 - \Sigma_A \end{pmatrix}.$$

If we take $I(A_i)$ to be the indicator that A_i is actually observed in the sample, and if we assume that the observation indicator is ignorable, then we can modify this to be

$$0 = \sum_{i=1}^{n} I(A_i) \begin{pmatrix} A_i - \mu_A \\ (A_i - \mu_A)^2 - \Sigma_A \end{pmatrix}.$$

As a result, we first note that for each mean, variance, and covariance associated with the terms in $\{X_1^*, \ldots, X_k^*, Z\}$, we can use these standard estimators. This will result in solving

the following system of equations,

$$0 = \sum_{i=1}^{n} \begin{pmatrix} I(X_{i1}^*)(X_{i1}^* - \mu_{X_1^*}) \\ I(X_{i1}^*)\left[(X_{i1}^* - \mu_{X_1^*})^2 - \Sigma_{X_1^*}\right] \\ \vdots \\ I(X_{ik}^*)(X_{ik}^* - \mu_{X_k^*}) \\ I(X_{ik}^*)\left[(X_{ik}^* - \mu_{X_k^*})^2 - \Sigma_{X_k^*}\right] \\ I(Z_i)(Z_i - \mu_Z) \\ I(Z_i)\left[(Z_i - \mu_Z)^2 - \Sigma_Z\right] \\ I(X_{i1}^*, X_{i2}^*)\left[(X_{i1}^* - \mu_{X_1^*})(X_{i2}^* - \mu_{X_2^*}) - \Sigma_{X_1^* X_2^*}\right] \\ \vdots \\ I(X_{i1}^*, X_{ik}^*)\left[(X_{i1}^* - \mu_{X_1^*})(X_{ik}^* - \mu_{X_k^*}) - \Sigma_{X_1^* X_k^*}\right] \\ I(X_{i1}^*, Z_i)\left[(X_{i1}^* - \mu_{X_1^*})(Z_i - \mu_Z) - \Sigma_{X_1^* Z}\right] \\ \vdots \\ I(X_{ik}^*, Z_i)\left[(X_{ik}^* - \mu_{X_k^*})(Z_i - \mu_Z) - \Sigma_{X_k^* Z}\right] \end{pmatrix}.$$

This leaves only the need to formulate the estimators surrounding the terms involving X, which are expressed as closed form estimators in Equation (3.5.2) and Equation (3.5.4) (when Z is observed), otherwise Equation (3.5.3). Note that none of these expressions depend on i, and are instead simply functions of the parameters listed above. As a result, we can simply use this previous expression with plug-in estimators to get the necessary results. If, however, we wish to jointly estimate these moment estimators as well, we can simply modify the above estimating equations to also include the closed form estimators.

First, note that the previous expression estimates a total of $(K+1)(2+K/2)$ terms. This includes $K+1$ terms for the mean, $K+1$ terms for the variances, and then $\frac{1}{2}(K+1)K$ covariance terms. Moreover, note that the closed form expressions of interest estimate 1 mean term, 1 covariance term (with Z), K terms for the multiplicative bias, and K covariance terms (with X_j^*), for a total of $2(K+1)$ terms. Use the notation $\mathbf{0}_\ell$ to denote a vector of zeros of size ℓ. Then, all of the moment parameters can be estimated by taking

$g_i(\xi)$ to be given by

$$\begin{pmatrix} I(X_{i1}^*)(X_{i1}^* - \mu_{X_1^*}) \\ I(X_{i1}^*)\left[(X_{i1}^* - \mu_{X_1^*})^2 - \Sigma_{X_1^*}\right] \\ \vdots \\ I(X_{iK}^*)(X_{iK}^* - \mu_{X_K^*}) \\ I(X_{iK}^*)\left[(X_{iK}^* - \mu_{X_K^*})^2 - \Sigma_{X_K^*}\right] \\ I(Z_i)(Z_i - \mu_Z) \\ I(Z_i)\left[(Z_i - \mu_Z)^2 - \Sigma_Z\right] \\ I(X_{i1}^*, X_{i2}^*)\left[(X_{i1}^* - \mu_{X_1^*})(X_{i2}^* - \mu_{X_2^*}) - \Sigma_{X_1^* X_2^*}\right] \\ \vdots \\ I(X_{i1}^*, X_{iK}^*)\left[(X_{i1}^* - \mu_{X_1^*})(X_{iK}^* - \mu_{X_K^*}) - \Sigma_{X_1^* X_K^*}\right] \\ I(X_{i1}^*, Z_i)\left[(X_{i1}^* - \mu_{X_1^*})(Z_i - \mu_Z) - \Sigma_{X_1^* Z}\right] \\ \vdots \\ I(X_{iK}^*, Z_i)\left[(X_{iK}^* - \mu_{X_K^*})(Z_i - \mu_Z) - \Sigma_{X_K^* Z}\right] \\ \mathbf{0}_{2(K+1)} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{(2+K/2)(K+1)} \\ \eta_{11} - \frac{\Sigma_{X_1^* Z}}{\Sigma_{XZ}} \\ \vdots \\ \eta_{1K} - \frac{\Sigma_{X_K^* Z}}{\Sigma_{XZ}} \\ \mu_X - \frac{1}{|J_0|}\sum_{j \in J_0} \frac{\mu_{X_j^*}}{\eta_{1j}} \\ \Sigma_{XZ} - \frac{1}{|J_1|}\sum_{j \in J_1} \Sigma_{X_j^* Z} \\ \Sigma_{XX_1^*} - \frac{1}{K-1}\sum_{\ell=2}^{K} \frac{\Sigma_{X_1^* X_\ell^*}}{\eta_{1\ell}} \\ \vdots \\ \Sigma_{XX_K^*} - \frac{1}{K-1}\sum_{\ell=1}^{K-1} \frac{\Sigma_{X_K^* X_\ell^*}}{\eta_{1\ell}} \end{pmatrix},$$

and then solving

$$0 = \sum_{i=1}^{n} g_i(\widehat{\xi}). \tag{B.1.1}$$

Note that if we do not have Z observed then we simply remove the terms referencing Z from the left component of g_i, and replace the right component with the estimators based on the other closed form expressions previously presented. This has assumed that we have scalar values X_i and Z_i. If instead these are vector valued, the form of the expression is exactly equivalent, however, for the variance and covariance terms we must vectorize the resulting expression (stacking the matrix components into a vector), and correspondingly update to the sizing of the zero vectors.

□

Proof of Lemma 3.5.2. Taking the definitions as stated in the Lemma, note that we have the parameter vector $\widehat{\Theta}_* = (\widehat{\Theta}', \widehat{\xi}')'$ solves the equation given by

$$0 = n^{-1}\sum_{i=1}^{n} \begin{bmatrix} U_n(\widehat{\Theta}, \widehat{\xi}) \\ g_n(\widehat{\xi}) \end{bmatrix},$$

and as a result we have

$$\sqrt{n}\left(\widehat{\Theta}_* - \Theta_*\right) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{A}^{-1}(\Theta,\xi)\,\mathcal{B}(\Theta,\xi)\,\mathcal{A}^{-1}(\Theta,\xi)'\right).$$

All that's left is then to note that $\sqrt{n}(\widehat{\Theta} - \Theta) = \sqrt{n}(Q\widehat{\Theta}_* - Q\Theta_*)$, and so a standard application of the Delta Method gives the necessary result. Note that the specific forms give

$$\begin{aligned}
A_{1,1}(\Theta,\xi) &= \frac{\partial}{\partial\Theta'}U_n(\Theta,\xi); & A_{1,2}(\Theta,\xi) &= \frac{\partial}{\partial\xi'}U_n(\Theta,\xi);\\
A_{2,1}(\Theta,\xi) &= \frac{\partial}{\partial\Theta'}g_n(\xi) = \mathbf{0}; & A_{2,2}(\Theta,\xi) &= \frac{\partial}{\partial\xi'}g_n(\xi);\\
B_{1,1}(\Theta,\xi) &= U_n(\Theta,\xi)U_n(\Theta,\xi)'; & B_{1,2}(\Theta,\xi) &= U_n(\Theta,\xi)g_n(\xi)';\\
B_{2,1}(\Theta,\xi) &= g_n(\xi)U_n(\Theta,\xi)'; & B_{2,2}(\Theta,\xi) &= g_n(\xi)g_n(\xi)';\\
\mathcal{A}(\Theta,\xi) &= E\left\{\begin{bmatrix}A_{1,1} & A_{1,2};\\ A_{2,1} & A_{2,2}\end{bmatrix}\right\}; & \mathcal{B}(\Theta,\xi) &= E\left\{\begin{bmatrix}B_{1,1} & B_{1,2}\\ B_{2,1} & B_{2,2}\end{bmatrix}\right\}.
\end{aligned}$$

□

Lemma B.1.1 (Conditional Means (extension of Lemma A.1 [10]))**.** *Assume that V_1, V_2, and V_3 are random vectors, and that $\delta > 0$ is a constant scalar. Take $E[V_3|V_2] = 0$, and denote $\mathrm{cov}(V_3|V_2 = v) = \Omega(v)$. Assume that $E[V_3|V_1]$ and $\mathrm{cov}(V_3|V_1)$ are three-times differentiable functions of δ, a.s. Then*

(a) If $V_1 = V_2 + \delta V_3$, then

$$E[V_3|V_1] = -\delta\left[\mathrm{Tr}\left\{\frac{\partial}{\partial v_1}\Omega(v_1)\right\} + \Omega(v_1)\frac{f'_{V_1}(v_1)}{f_{V_1}(v_1)}\right]_{v_1=V_1} + O_p(\delta^2), \tag{B.1.2}$$

and

$$\mathrm{cov}(V_3|V_1) = \Omega(V_1) + O_p(\delta). \tag{B.1.3}$$

(b) If $V_1 = V_2(\mathbf{1} + \delta V_3)$, then

$$E[V_3|V_1] = -\delta\left[\mathrm{diag}\left\{\Omega(v_1)\right\} + v_1 \circ \mathrm{Tr}\left\{\frac{\partial}{\partial v_1}\Omega(v_1)\right\} + v_1 \circ \Omega(v_1)\frac{f'_{V_1}(v_1)}{f_{V_1}(v_1)}\right]_{v_1=V_1} + O_p(\delta^2), \tag{B.1.4}$$

and

$$\mathrm{cov}(V_3|V_1) = \Omega(V_1) + O_p(\delta). \tag{B.1.5}$$

Proof for Lemma B.1.1. While (a) was demonstrated in the proof for Lemma A.1 [10], we include the full detail here as it is instructive for proving (b).

First note that, when $\delta = 0$ we get $V_1 = V_2$ in both (a) and (b). As a result, for $\delta \approx 0$, an arbitrary function of V_1, $h_{V_1}(v)$ is such that $h_{V_1}(v) = h_{V_2}(v) + O_p(\delta)$, from a first-order Taylor expansion. This gives the results for both covariance terms. It also gives the fact that, in either scenario, we can write $f'_{V_2}(v)/f_{V_2}(v) = f'_{V_1}(v)/f_{V_1}(v) + O_p(\delta)$. Then,

$$E\left[V_3 \middle| V_1\right] = \frac{1}{f_{V_1}(V_1)} \int v_3 f_{V_1,V_3}(V_1, v_3) dv_3$$
$$\begin{cases} \overset{\text{(a)}}{=} \frac{1}{f_{V_1}(V_1)} \int v_3 f_{V_3|V_2}(v_3|V_2 = V_1 - \delta v_3) f_{V_2}(V_1 - \delta v_3) dv_3, \\ \overset{\text{(b)}}{=} \frac{1}{f_{V_1}(V_1)} \int v_3 (\mathbf{1} + \delta v_3)^{-1} f_{V_3|V_2}(v_3|V_2 = V_1(\mathbf{1} + \delta v_3)^{-1}) f_{V_2}(V_1(\mathbf{1} + \delta v_3)^{-1}) dv_3. \end{cases}$$

The remainder of the proof follows by considering Taylor expansions of the integrands around $\delta = 0$, and noting that $f_{V_1}(V_1) = f_{V_2}(V_1) + O_p(\delta)$. Taking first the expression for (a), note that evaluating the expression at $\delta = 0$ gives $v_3 f_{V_3|V_2}(v_3|V_2 = V_1) f_{V_2}(V_1)$, which integrating gives $E[V_3|V_2 = V_1] f_{V_2}(V_1) = 0$ by assumption. Differentiating the integrand, and evaluating at $\delta = 0$ gives the expression

$$-v_3 v_3' f'_{V_3|V_2}(v_3|V_2 = V_1) f_{V_2}(V_1) - v_3 v_3' f_{V_3|V_2}(v_3|V_2 = V_1) f'_{V_2}(V_1),$$

where the prime on the conditional density represents the derivative with respect to the conditioning term. Integrating these terms gives

$$-\left[\mathrm{Tr}\left\{\frac{\partial}{\partial v_1}\Omega(v_1)\right\} f_{V_2}(V_1) + \Omega(V_1) f'_{V_2}(v_1)\right]_{v_1 = V_1}.$$

Combining this with the Taylor expansion in the denominator gives the desired result.

For (b) we follow a similar strategy. The integral evaluates to 0 when $\delta = 0$ (by assumption), and the first derivative of the integrand with $\delta = 0$ is given by

$$-v_3 \circ v_3 f_{V_3|V_2}(V_3|V_2 = V_1) f_{V_2}(V_1) - v_3 v_3' f'_{V_3|V_2}(V_3|V_2 = V_1) - v_3 v_3' f_{V_3|V_2}(V_3|V_2 = V_1) f'_{V_2}(V_1).$$

Once again, we integrate giving

$$-\left[\mathrm{diag}\left\{\Omega(v_1)\right\} f_{V_2}(V_1) + \mathrm{Tr}\left\{\frac{\partial}{\partial v_1}\Omega(v_1)\right\} f_{V_2}(V_1) + \Omega(V_1) f'_{V_2}(v_1)\right]_{v_1 = V_1}.$$

Then expanding the denominator as with (a) gives us the necessary result. □

Proof of Theorem 3.6.1. This theorem follows as a direct application of Lemma B.1.1. For the additive case, we consider $V_1 \equiv X^*$, $V_2 \equiv \eta_0 + \eta_1 X$, and $V_3 \equiv U$. Then, it is clear that $E[U|\eta_0 + \eta_1 X] = 0$, by our outlined assumptions, and as a result, $E[U|X^*] = -\delta \left[\text{Tr}\left(\frac{\partial \Omega(x)}{\partial x} \right) + \Omega(x) \frac{f'_{X^*}(x)}{f_{X^*}(x)} \right]_{x=X^*} + O_p(\delta^2)$. Now, since $X = \eta_1^{-1}(X^* - \eta_0 - \delta U)$, the results follows directly. The multiplicative case requires additional considerations, but is otherwise similar.

First, taking $V_1 \equiv X^* - \eta_0$, $V_2 \equiv \eta_1 X$, and $V_3 \equiv U$, then we note that $E[U|X^* = x] = E[V_3|V_1 = x - \eta_0]$ and $\text{cov}(U|X^* = x) = \text{cov}(V_3|V_1 = x - \eta_0)$. Additionally, $f_{V_1}(v) = f_{X^*}(v + \eta_0)$. Then, in order to solve for $E[X|X^*]$, we make use of a Taylor expansion of $X = (\mathbf{1} + \delta U)^{-1}(X^* - \eta_0)$, around $\delta = 0$, to handle the ratio. We consider the second order expansion so as to maintain an error of order $O_p(\delta^3)$ overall. That is, consider

$$
\begin{aligned}
(1 + \delta U)^{-1} &= 1 - \delta U + \delta^2 \text{diag}\,(UU') + O_p(\delta^3) \\
&\Longrightarrow E\left[(1 + \delta U)^{-1} \middle| X^* \right] \\
&\quad = 1 - \delta E[U|X^*] + \delta^2 \text{diag}\,(\text{cov}(U|X^*) + E[U|X^*]E[U|X^*]') + O_p(\delta^3) \\
&\quad = 1 - \delta E[U|X^*] + \delta^2 \text{diag}\,(\text{cov}(U|X^*)) + O_p(\delta^3),
\end{aligned}
$$

where the last equality holds since $\delta^2 E[U|X^*]E[U|X^*]' = O_p(\delta^4)$. Then, noting that

$$
\begin{aligned}
&E[U|X^* = x] \\
&= E[V_3|V_1 = x - \eta_0] \\
&= -\delta \left[\text{diag}\left\{ \Omega(v_1) \right\} + v_1 \circ \text{Tr}\left\{ \frac{\partial}{\partial v_1} \Omega(v_1) \right\} + v_1 \circ \Omega(v_1) \frac{f'_{V_1}(v_1)}{f_{V_1}(v_1)} \right]_{v_1 = x - \eta_0} + O_p(\delta^2) \\
&= -\delta \left[\text{diag}\left\{ \Omega(x - \eta_0) \right\} + (x - \eta_0) \circ \left\{ \text{Tr}\left(\frac{\partial \Omega(v)}{\partial v} \right) \Big|_{v = x - \eta_0} + \Omega(x - \eta_0) \frac{f'_{X^*}(x)}{f_{X^*}(x)} \right\} \right] \\
&+ O_p(\delta^2).
\end{aligned}
$$

Combining these two quantities gives the desired result. □

Proof of Theorem 3.6.2. The proof will be presented, where convenient, using notation that implies scalar X. This can be extended to the multivariate case by carefully vectorizing the relevant M-estimators.

First, note that $\widehat{\Theta}_{\text{RC}}$ solves $U_n(Y, Z, \widehat{X}, \widehat{\Theta}_{\text{RC}}) = 0$. Now, $\widehat{X} = \widehat{\mu} + \widehat{\beta}\sum_{j=1}^{k}\widehat{\alpha}_k X^* + \widehat{\gamma}Z$ where the estimators $(\widehat{\mu}, \widehat{\beta}, \widehat{\gamma}, \{\widehat{\alpha}_j\}_j)$ solve

$$h\left(\widehat{\mu}, \widehat{\beta}, \widehat{\gamma}, \{\widehat{\alpha}_j\}_j\right) = \begin{pmatrix} \mu_X - \widehat{\mu} - \widehat{\beta}\mu_{X^*} - \widehat{\gamma}\mu_Z \\ \Sigma_{XX^*} - \widehat{\mu}\mu_{X^*} - \widehat{\beta}\Sigma_{X^*X^*} - \widehat{\gamma}\Sigma_{ZX^*} \\ \Sigma_{XZ} - \widehat{\mu}\mu_Z - \widehat{\beta}\Sigma_{X^*Z} - \widehat{\gamma}\Sigma_{ZZ} \\ \text{Tr}\left\{\Sigma_{XX_j^*} - \widehat{\beta}\Sigma_{X^*X_j^*} - \widehat{\gamma}\Sigma_{ZX_j^*}\right\}_{j=1}^{k} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \{0\}_{j=1}^{k} \end{pmatrix}.$$

Here, the reliance of the first three components on α_j is suppressed in X^*. By Lemma 3.5.1 each of these components are estimable using an M-estimator. The previous results frame μ_j, $\Sigma_{X_j^*X_l^*}$, $\Sigma_{XX_j^*}$, and $\Sigma_{X_j^*Z}$ in place of μ_{X^*}, $\Sigma_{X^*X^*}$, Σ_{XX^*}, Σ_{ZX^*}, and $\Sigma_{X^*X_j^*}$. However, the latter can be written as transformations of the former. Specifically,

$$\mu_{X^*} = \sum_{j=1}^{k}\alpha_j\mu_j; \qquad \Sigma_{X^*X^*} = \sum_{j=1}^{k}\sum_{l=1}^{k}\alpha_j\alpha_l\Sigma_{X_j^*X_l^*}; \qquad \Sigma_{XX^*} = \sum_{j=1}^{k}\alpha_j\Sigma_{XX_j^*};$$
$$\Sigma_{X^*Z} = \sum_{j=1}^{k}\alpha_j\Sigma_{X_j^*Z}; \qquad \Sigma_{X^*X_j^*} = \sum_{l=1}^{k}\alpha_l\Sigma_{X_l^*X_j^*}.$$

Now, noting that $h\left(\widehat{\mu}, \widehat{\beta}, \widehat{\gamma}, \widehat{\alpha}\right) = \mathbf{0} \iff n^{-1}\sum_{i=1}^{n} h\left(\widehat{\mu}, \widehat{\beta}, \widehat{\gamma}, \widehat{\alpha}\right) = \mathbf{0}$, this means that we can stack $g_i(\cdot)$ with $h(\cdot)$ which forms an estimating equation for the relevant parameters. Then, this can be stacked with $\Psi(\cdot)$, as the estimator of Θ_{RC} is given as solution to $n^{-1}\sum_{i=1}^{n}\Psi(Y_i, Z_i, \widehat{\mu} + \widehat{\beta}X_i^* + \widehat{\gamma}Z_i, \widehat{\Theta}_{\text{RC}}) = 0$. As a result, the asymptotic distribution of $\widehat{\Theta}_{\text{RC}}$ can be derived through the standard theory, using the M-estimator

$$n^{-1}\sum_{i=1}^{n}\begin{pmatrix} \Psi(Y_i, Z_i, \widehat{\mu} + \widehat{\beta}X_i^*(\widehat{\alpha}) + \widehat{\gamma}Z_i, \widehat{\Theta}_{\text{RC}}) \\ h(\widehat{\mu}, \widehat{\beta}, \widehat{\gamma}, \widehat{\xi}, \widehat{\alpha}) \\ g_i(\widehat{\xi}) \end{pmatrix} = \mathbf{0}.$$

For the first matrix in the asymptotic covariance, denoted $\mathcal{A}_{\text{RC}}$, defined as the expectation of the 3×3 block matrix given by the derivatives of the previous estimating equation. We note that this will be an upper triangular matrix since h is independent of Θ, and g_i is independent of Θ and $(\mu, \beta, \gamma, \alpha)$. Of course the precise form of this matrix will rely on

the estimating equation for Θ, and on the data available defining g_i. Generally,

$$\begin{aligned}\mathcal{A}_{\text{RC}} &= \begin{bmatrix} \mathcal{A}_{RC}^{(1,1)} & \mathcal{A}_{RC}^{(1,2)} & \mathcal{A}_{RC}^{(1,3)} \\ \mathbf{0} & \mathcal{A}_{RC}^{(2,3)} & \mathcal{A}_{RC}^{(2,3)} \\ \mathbf{0} & \mathbf{0} & \mathcal{A}_{RC}^{(3,3)} \end{bmatrix} \\ &= \begin{bmatrix} E\left[\Psi_\Theta(\Theta,\mu,\beta,\gamma,\alpha,\xi)\right] & E\left[\Psi_{(\mu,\beta,\gamma,\alpha)}(\Theta,\mu,\beta,\gamma,\alpha,\xi)\right] & E\left[\Psi_\xi(\Theta,\mu,\beta,\gamma,\alpha,\xi)\right] \\ \mathbf{0} & E\left[h_{(\mu,\beta,\gamma,\alpha)}(\mu,\beta,\gamma,\alpha,\xi)\right] & E\left[h_\xi(\mu,\beta,\gamma,\alpha,\xi)\right] \\ \mathbf{0} & \mathbf{0} & E\left[g_\xi(\xi)\right] \end{bmatrix},\end{aligned}$$

with $W_\Delta(\Delta)$ representing the derivative of $W(\cdot)$ with respect to Δ'. Similarly,

$$\mathcal{B}_{\text{RC}} = \begin{bmatrix} B_{\text{RC}}^{(1,1)} & B_{\text{RC}}^{(1,2)} & B_{\text{RC}}^{(1,3)} \\ B_{\text{RC}}^{(2,1)} & B_{\text{RC}}^{(2,2)} & B_{\text{RC}}^{(1,3)} \\ B_{\text{RC}}^{(3,1)} & B_{\text{RC}}^{(3,2)} & B_{\text{RC}}^{(3,3)} \end{bmatrix} = \begin{bmatrix} B_{\text{RC}}^{(1,1)} & B_{\text{RC}}^{(1,2)} & B_{\text{RC}}^{(1,3)} \\ B_{\text{RC}}^{(2,1)} & B_{\text{RC}}^{(2,2)} & B_{\text{RC}}^{(1,3)} \\ B_{\text{RC}}^{(3,1)} & B_{\text{RC}}^{(3,2)} & B_{\text{RC}}^{(3,3)} \end{bmatrix} = \begin{bmatrix} E[\Psi\Psi'] & \mathbf{0} & E[\Psi g'] \\ \mathbf{0} & E[hh'] & \mathbf{0} \\ E[g\Psi'] & \mathbf{0} & E[gg'] \end{bmatrix},$$

where the zeros come from noting that, since $E[\Psi] = E[g] = \mathbf{0}$, and that h is constant (with respect to the underlying random variables), we have that $E[\Psi h'] = E[gh'] = \mathbf{0}$. Note that, in fact, the structure of g is such that many of the components in the top right (and by symmetry bottom left) will also have this zero property, though, upon specification of g this should become obvious. The standard theory of M-estimators then gives the asymptotic covariance of the stacked estimator as $\mathcal{A}_{\text{RC}}^{-1}\,\mathcal{B}_{\text{RC}}\,\mathcal{A}_{\text{RC}}^{-'}$. □

Proof of Theorem 3.7.1. The two proposed estimators for the SIMEX correction – whether averaged before or after extrapolation – can have their asymptotic distribution derived as an extension of [8] and Lemma 3.5.2. Our primary interest lies in $\widehat{\Theta} = \mathcal{G}(-1, \widehat{\Gamma})$, where $\widehat{\Gamma}$ is the parameter vector that minimizes $R(\Gamma)'C^{-1}R(\Gamma)$. Here, C is a positive-definite matrix, decided on by the analyst (for instance, $C = I$ for standard least squares), $R(\Gamma) = \widehat{\Theta}_\Lambda - \mathcal{G}(\Lambda, \Gamma)$, and $\widehat{\Theta}_\Lambda$ is the vector formed by stacking $(\widehat{\Theta}_{\lambda_1}, \ldots, \widehat{\Theta}_{\lambda_R})$. $\widehat{\Theta}_\lambda$ is given by $B^{-1}\sum_{b=1}^{B}\widehat{\Theta}_{b,\lambda}$ for each $\lambda \in \Lambda$, and $\widehat{\Theta}_{b,\lambda}$ solves $n^{-1}\sum_{i=1}^{n}\psi(Y_i, Z_i, X_{bi}^*(\lambda), \Theta_\lambda) = 0$. Thus, we work to derive the asymptotic distribution of $\sqrt{n}(\widehat{\Gamma} - \Gamma)$, and then apply the Delta method for the necessary results.

Note that, by definition we have $\Theta_\Lambda = \mathcal{G}(\Gamma, \Lambda)$ and $\widehat{\Theta}_\Lambda = \mathcal{G}(\widehat{\Gamma}, \Lambda)$. Define $s(\Gamma) = \frac{\partial}{\partial\Gamma}\mathcal{G}(\Gamma, \Lambda)'$. A Taylor expansion of $\mathcal{G}$ results in $\mathcal{G}(\widehat{\Gamma}, \Lambda) = \mathcal{G}(\Gamma, \Lambda) + s(\Gamma)'\left\{\widehat{\Gamma} - \Gamma\right\} + o_p(1)$, which re-arranging and multiplying by $\sqrt{n}s(\Gamma)C^{-1}$ (for invertibility), and defining $\Omega(\Gamma) =$

$s(\Gamma)C^{-1}s(\Gamma)'$ gives that

$$\sqrt{n}\left(\widehat{\Gamma}-\Gamma\right)=\Omega(\Gamma)^{-1}s(\Gamma)C^{-1}\cdot\sqrt{n}\left(\widehat{\Theta}_{\Lambda}-\Theta_{\Lambda}\right)+o_p(1).$$

As a result, we can focus the proof on the asymptotic distribution of $\sqrt{n}\left(\widehat{\Theta}_{\Lambda}-\Theta_{\Lambda}\right)$, and then apply a straightforward transformation for the distribution of $\widehat{\Gamma}$.

For the estimator computed as the average after extrapolation, we focus on $\widehat{\Theta}^{(j)}_{\text{SIMEX}}(\lambda)$, which use Equation 3.7.1 directly for error term j. Stacking each of these estimators over the values of $\lambda\in\Lambda$, we get $\widehat{\Theta}^{(j)}_{\Lambda}$, and then consider the stacked version, stacking over $j=1,\ldots,k$, to be given by $\widehat{\Theta}_{\Lambda}$. This notation must be extended to the other relevant parameters: Γ_j for the j-th extrapolant values, giving $\Omega(\Gamma_j)=s_j(\Gamma_j)C_j^{-1}s_j(\Gamma_j)'$. Then, the transformations here apply for each j, and so $\Omega(\Gamma)$, $s(\Gamma)$, and C are formed by taking the block diagonal matrices over all j. With these amendments, the following argument applies directly. Once joint estimators are obtained for each $\mathcal{G}(-1,\Gamma_j)$, the final distribution can be taken by applying the relevant averaging transformation.

An asymptotic linearization of $n^{-1}\sum_{i=1}^{n}\psi(Y_i,Z_i,X^*_{bi}(\lambda),\Theta_\lambda)=0$ leads to

$$\sqrt{n}\left(\widehat{\Theta}_{b,\lambda}-\Theta_\lambda\right)=\mathcal{A}^{(1,1)^{-1}}(\lambda)\sqrt{n}\sum_{i=1}^{n}\psi(Y_i,Z_i,X^*_{bi}(\lambda),\Theta_\lambda)+o_p(1),$$

where $\mathcal{A}^{(1,1)^{-1}}(\lambda)=E\left[\frac{\partial}{\partial\Theta'}\psi(Y,Z,X^*_b(\lambda),\Theta_\lambda)\right]$. Then, averaging both sides over b, results in $\sqrt{n}\left(\widehat{\Theta}_\lambda-\Theta_\lambda\right)=\mathcal{A}^{(1,1)^{-1}}(\lambda)\sqrt{n}\sum_{i=1}^{n}B^{-1}\sum_{b=1}^{B}\psi(Y_i,Z_i,X^*_{bi}(\lambda),\Theta_\lambda)+o_p(1)$. This result holds for all $\lambda\in\Lambda$, where Λ is taken to be the fixed grid of size R that we simulate at.

The computation of these estimators, however, rely on the components of ξ identified in Lemma 3.5.1, through the estimating equation $n^{-1}\sum_{i=1}^{n}g_i(\cdot)=0$, and on the weights α. We specify an M-estimator for each α_j, based on some optimality criteria, and include the weights α_j in ξ. For both estimators under consideration, all parameters required for correction are then contained in ξ, and we can write

$$n^{-1}\sum_{i=1}^{n}\psi(Y_i,Z_i,X^*_{bi}(\lambda),\Theta_\lambda)=n^{-1}\sum_{i=1}^{n}\psi\left(Y_i,Z_i,\eta_{1\cdot}^{-1}\circ\left[X^*_i-\eta_{0\cdot}+\sqrt{\lambda}M_*^{1/2}\nu_b\right],\Theta_\lambda\right),$$

which we define to be $n^{-1}\sum_{i=1}^{n}\psi_{ib}(\lambda)$, where the necessary alterations are made to have this stacked over j as discussed above. Writing the joint M-estimator as $n^{-1}\sum_{i=1}^{n}\begin{bmatrix}\psi_{ib}(\lambda)\\ g_i\end{bmatrix}=0$,

and applying the exact argument as above, we get that

$$\sqrt{n}\left(\begin{bmatrix}\widehat{\Theta}_\lambda \\ \widehat{\xi}\end{bmatrix} - \begin{bmatrix}\Theta_\lambda \\ \xi\end{bmatrix}\right) = \mathcal{A}^{-1}(\lambda)\sqrt{n}\sum_{i=1}^{n} B^{-1}\sum_{b=1}^{B}\begin{bmatrix}\psi_{ib}(\lambda) \\ g_i\end{bmatrix} + o_p(1)$$
$$:= \mathcal{A}^{-1}(\lambda)\sqrt{n}\sum_{i=1}^{n}\begin{bmatrix}\Psi_i(\lambda) \\ g_i\end{bmatrix} + o_p(1),$$

where, again, the last equality is taken to be a notational definition. We have that

$$\mathcal{A}(\lambda) = \begin{bmatrix}A^{(1,1)} & A^{(1,2)} \\ \mathbf{0} & A^{(2,2)}\end{bmatrix} = \begin{bmatrix}E\left[\frac{\partial}{\partial\Theta'}\psi(Y,Z,X_b^*(\lambda),\Theta_\lambda)\right] & E\left[\frac{\partial}{\partial\xi'}\psi(Y,Z,X_b^*(\lambda),\Theta_\lambda)\right] \\ \mathbf{0} & E\left[\frac{\partial}{\partial\xi'}g(\xi)\right]\end{bmatrix}.$$

Define Θ_Λ to be the vector stacking $(\Theta_{\lambda_1}, \Theta_{\lambda_2}, \ldots, \Theta_{\lambda_R})$, with the corresponding definition for $\widehat{\Theta}_\Lambda$, $\widetilde{\Psi}_i(\Lambda)$ to be the vector stacking $(\Psi_i(\lambda_1), \ldots, \Psi_i(\lambda_R), g_i)$, and $\mathcal{A}_{\text{SIMEX}}(\Lambda)$ to be the matrix with $\mathcal{A}^{(1,1)}(\lambda_1), \mathcal{A}^{(1,1)}(\lambda_2), \ldots, \mathcal{A}^{(1,1)}(\lambda_R)$ on the diagonals first R diagonals, and then an $R+1$ column with $(\mathcal{A}^{(1,2)}(\lambda_1), \ldots, \mathcal{A}^{(1,2)}(\lambda_R), \mathcal{A}^{(2,2)})$, then zeros elsewhere. Note that the $\mathcal{A}^{(2,2)}$ portion of the matrix is constant across all λ, and so this matrix forms a block upper triangular matrix, with $(R+1)\times(R+1)$ blocks; each row j takes the relevant matrix from $\mathcal{A}(\lambda_j)$ in the j-th block, and takes the cross matrix in the $R+1$ block. The above result implies that

$$\sqrt{n}\left(\begin{bmatrix}\widehat{\Theta}_\Lambda \\ \widehat{\xi}\end{bmatrix} - \begin{bmatrix}\Theta_\Lambda \\ \xi\end{bmatrix}\right) = \mathcal{A}^{-1}(\lambda)\sqrt{n}\sum_{i=1}^{n}\widetilde{\Psi}_i(\Lambda) + o_p(1).$$

Standard asymptotic theory then gives that this converges in distribution to a mean zero normal distribution, with variance given by $\mathcal{A}^{-1}(\Lambda)\Sigma\,\mathcal{A}^{-1}(\Lambda)'$, where $\Sigma = E\left[\widetilde{\Psi}(\Lambda)\widetilde{\Psi}(\Lambda)'\right]$. To extract only the distribution of $\sqrt{n}\left(\widehat{\Theta}_\Lambda - \Theta_\Lambda\right)$, we multiply by

$$Q = \begin{bmatrix}I_{\dim\Theta_\Lambda\times\dim\Theta_\Lambda} & \mathbf{0}_{\dim\Theta_\lambda\times\dim\xi}\end{bmatrix}$$

, giving the same mean zero with covariance $Q\,\mathcal{A}^{-1}(\Lambda)\Sigma\,\mathcal{A}^{-1}(\Lambda)'Q'$. Combining this with the previous discussion gives

$$\sqrt{n}\left(\widehat{\Gamma} - \Gamma\right) \xrightarrow{d} N\left(0, \Sigma_*\right),$$

where $\Sigma_* = \Omega^{-1}(\Gamma)s(\Gamma)C^{-1}Q\,\mathcal{A}^{-1}(\Lambda)\Sigma\,\mathcal{A}^{-1}(\Lambda)'Q'C^{-1'}s(\Gamma)'\Omega^{-1}(\Gamma)'$.

Finally, the SIMEX estimators are defined by taking the estimated $\widehat{\Gamma}$, and plugging into $\mathcal{G}$ at $\lambda = -1$. As a result, we apply the Delta method with $\mathcal{G}(-1,\cdot)$ as the function, and (assuming that it satisfies the requisite properties) we find that

$$\sqrt{n}\left(\mathcal{G}\left(\widehat{\Gamma},-1\right) - \mathcal{G}\left(\Gamma,-1\right)\right) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{G}_\Gamma(-1,\Gamma)\Sigma_*\,\mathcal{G}_\Gamma(-1,\Gamma)'\right).$$

If using the estimator which has been averaged prior to extrapolation, this gives us the required distribution. Otherwise, this has resulted in a $(k\dim\Theta)\times 1$ stacked estimator $\mathcal{G}(-1,\widehat{\Gamma})$, and so $\widehat{\Theta}_{\text{SIMEX}}$ is given by multiplying through the matrix Q^* which is given by $\begin{bmatrix}\alpha_1 I_{\dim\Theta} & \cdots & \alpha_k I_{\dim\Theta}\end{bmatrix}$, where $\sum_{j=1}^k \alpha_j = 1$. This results in a final asymptotic covariance of $Q^*\,\mathcal{G}_\Gamma(-1,\Gamma)\Sigma_*\,\mathcal{G}_\Gamma(-1,\Gamma)'Q^{*'}$.

While the notational conventions were the same for either the averaging before, or the averaging afterwards, we note that the matrix structures are fundamentally different between the two. This is true even before the adjustment with Q^*, since $\Omega(\Gamma)$, $s(\Gamma)$, C, Q, $\mathcal{A}$ and Σ are all of different forms and shapes. □

B.2 Chapter 4

Proof of Lemma 4.4.1. Consider $\varphi_{U_\lambda}(t) = \varphi_U(t)\exp\left(-\frac{\lambda t^2\sigma^2}{2}\right)$. The second order Taylor expansion of this is given by

$$\varphi_{U_\lambda}(t) \approx 1 + t\varphi^{(1)}_{U_\lambda}(0) + \frac{t^2}{2}\varphi^{(2)}_{U_\lambda}(0) = 1 - \frac{t^2}{2}(1+\lambda)\sigma^2.$$

Here, we've used the fact that $E[U] = 0$ and $E[U^2] = -\varphi^{(2)}_U(0)$. Taking $\lambda = -1$ is the unique solution that makes the second order approximation exactly 1. □

Proof of Theorem 4.4.2. First note that since U is symmetric, with variance σ^2, then $\varphi_U(t)$ is a real-valued function, with $\varphi^{(1)}_U(0) = 0$ and $\varphi^{(2)}_U(0) = -\sigma^2$. We wish to know when

$$\left(1 - \varphi_U(t)\exp\left(\frac{t^2\sigma^2}{2}\right)\right)^2 \leq \left(1 - \varphi_U(t)\exp\left(-\lambda\frac{t^2\sigma^2}{2}\right)\right)^2$$
$$\iff 0 \leq \varphi_U(t)^2\left(\exp\left(-\lambda t^2\sigma^2\right) - \exp\left(t^2\sigma^2\right)\right) - 2\varphi_U(t)\left(\exp\left(-\lambda\frac{t^2\sigma^2}{2}\right) - \exp\left(\frac{t^2\sigma^2}{2}\right)\right).$$

For all $\lambda > -1$, this condition becomes $0 \geq \mathcal{H}(t, \lambda)$, and for otherwise $0 \leq \mathcal{H}(t, \lambda)$, where $\mathcal{H}(t, \lambda)$ is given by $\varphi_U(t) \left[\varphi_U(t) \left(\frac{\exp\left(-\frac{\lambda t^2 \sigma^2}{2}\right) + \exp\left(\frac{t^2 \sigma^2}{2}\right)}{2} \right) - 1 \right]$.

Note that $\mathcal{H}(0, \lambda) = 0$, and that $\mathcal{H}'(0, \lambda) = 0$. Further, $\mathcal{H}''(0, \lambda) = \frac{-\sigma^2 - \lambda \sigma^2}{2}$. For $\lambda > -1$, this will take negative values and will be positive otherwise. This means that $t = 0$ will give a local maximum for $\mathcal{H}(t, \lambda)$ if $\lambda > -1$ and a local minimum if $\lambda < -1$. Invoking continuity of $\mathcal{H}(t, \lambda)$, this must mean that there is a region around $t = 0$ for which $0 \geq \mathcal{H}(t, \lambda)$ when $\lambda > -1$ and $0 \leq \mathcal{H}(t, \lambda)$ otherwise. □

Proof of Theorem 4.5.1. Let $\widehat{F}_\lambda$ represent the distribution function for our resampled covariates, $X_i^* + \sum_{j=1}^{\lambda} U_{ij}^*$, and F_λ be the distribution function of $X_i + \sum_{j=0}^{\lambda} \widetilde{U}_{ij}$. If we take $*$ to represent the convolution operator, then for independent X and W we have that the distribution function of $X + W$ is given by $F_X * F_W$. Similarly, for independent $W_1, \ldots, W_n$, we can write the distribution function for $\sum_{\ell=1}^{n} W_\ell$ as $F_W * F_W * \cdots * F_W$, with n terms, which we denote $F_W^{(*n)}$. With this notation, we have $\widehat{F}_\lambda = F_X * F_{\widetilde{U}} * F_{U^*}^{(*\lambda)} = F_X * F_{\widetilde{U}} * \widehat{F}_{\widetilde{U}}^{(*\lambda)}$, and the distribution function $F_\lambda = F_X * F_{\widetilde{U}}^{(*(\lambda+1))}$. The convolution of F_X and $\widehat{F}_W$ converges almost surely to the convolution between F_X and F_W, since

$$
\begin{aligned}
&P\left\{ \lim_{n \to \infty} \int_0^{\mathcal{T}} F_X(\tau) \left[\widehat{F}_W(t - \tau) - F_W(t - \tau) \right] d\tau = 0 \right\} \\
&\geq P\left\{ \lim_{n \to \infty} \widehat{F}_W(\tau) = F_W(\tau) \quad \forall\ \tau \in [0, t] \right\} = 1.
\end{aligned}
$$

This follows through dominated convergence, bringing the limit through the integral, then noting that the set of events where the difference in convolutions (left-hand side) is zero is a superset of the set of events where the empirical distribution function equals the true distribution function (right-hand side). The right-hand side is one since the empirical CDF converges almost surely.

This result gives us almost sure convergence of $\widehat{F}_\lambda$ to F_λ. Moreover, the characteristic function of F_λ is given by $\varphi_X(t) \varphi_{\widetilde{U}}(t)^{\lambda+1}$, and so as $\lambda \to -1$ we have $F_\lambda \longrightarrow F_X$. Taking the functional representation we have $\widehat{\theta}_\lambda = \mathbf{T}(\widehat{F}_\lambda) \longrightarrow \mathbf{T}(F_\lambda) = \mathcal{G}(\lambda)$ by weak continuity, which states that if F_n converges to F, then $\mathbf{T}(F_n)$ converges almost surely to $\mathbf{T}(F)$. Then, since this is true for all λ, and since when $\lambda = -1$, $F_\lambda = F_X$, the correct specification and consistent estimation of $\mathcal{G}$ gives the result. □

Proof of Theorem 4.5.2. The regularity conditions on the functional $\mathbf{T}$ are such that a linearization can be obtained. Specifically, we should find that, $n^{1/2}(\mathbf{T}(\widehat{F}_\lambda) - \mathbf{T}(F_\lambda)) =$

$n^{-1/2}\sum_{i=1}^{n}\psi_{F_\lambda,i} + o_p(1)$, where ψ_{F_λ} represents the influence curve of the functional at F_λ. This representation is sufficient to continue to the proof from Carroll, Küchenhoff, Lombard, and Stefanski [8], verbatim, as it is equivalent to the M-estimator representation that they derive for $\widehat{\theta}_\lambda$. □

B.3 Chapter 6

Proof of Theorem 6.3.1. First note that based on the (approximate) non-differential error, and the linear form for the blip, we have

$$\begin{aligned} E[Y|H_K^*, A_K] &= E[E[Y|H_K, H_K^*, A_K]|H_K^*, A_K] \\ &= E[f_K(H_K)|H_K^*, A_K] + A_K E[H_K|H_K^*]'\psi_K \\ &= E[f_K(H_K)|H_K^*] + A_K \widehat{H}_K'\psi_K. \end{aligned}$$

Moreover, note that taking $(\widehat{\beta}_K, \widehat{\psi}_K)$ to be the WLS estimators, then under standard regularity conditions $(\widehat{\beta}_K, \widehat{\psi}_K)$ will be consistent for (β_K^*, ψ_K^*) which are the (unique) solution to $E[U_K(\beta_K^*, \psi_K^*)] = 0$, where

$$U_K(\beta_K^*, \psi_K^*) = \begin{pmatrix} \widehat{H}_K \\ A_K\widehat{H}_K \end{pmatrix} w_K(A_K, \widehat{H}_K)\left\{Y - \widehat{H}_K'\beta_K^* - A_K\widehat{H}_K'\psi_K^*\right\}. \tag{B.3.1}$$

We will show that, supposing (β_K^*, ψ_K^*) are unique, then under either (A1) or (A2) we must have $\psi_K^* = \psi_K$.

If (A1) holds then we have that $E[Y|H_K^*, A_K] = \widehat{H}_K'\beta_K + A_K\widehat{H}_K'\psi_K$. Then, from Equation(B.3.1) we see that

$$E\left[U_K(\beta_K^*, \psi_K^*)|H_K^*, A_K\right] = \begin{pmatrix} \widehat{H}_K \\ A_K\widehat{H}_K \end{pmatrix} w_K(A_K, \widehat{H}_K)\left\{\widehat{H}_K'\left[\beta_K - \beta_K^*\right] + A_K\widehat{H}_K'\left[\psi_K - \psi_K^*\right]\right\}.$$

The uniqueness of the root demonstrates that at $\beta_K^* = \beta_K$ and $\psi_K^* = \psi_K$ we have $E[U_K|H_K^*, A_K] = 0$, which gives consistency of $\widehat{\psi}_K$ for ψ_K, as needed.

If (A2) holds then, through the use of two-step M-estimation techniques we can replace $\pi_K(\widehat{H}_K)$ with $P(A_K = 1|\widehat{H}_K)$ in the asymptotic analysis. Doing so (maintaining the notation of $\pi_K(\cdot)$), we can consider the two sets of equations implied by Equation(B.3.1)

separately.

$$
\begin{aligned}
&E[U_K^{(2)}(\beta_K^*, \psi_K^*)|H_K^*] \\
&= E\left[A_K \widehat{H}_K w_K(A_K, \widehat{H}_K)\left\{E[f(H_K)|H_K^*] - \widehat{H}_K'\beta_K^* + A_K\widehat{H}_K'\left[\psi_K - \psi_K^*\right]\right\}\Big| H_K^*\right] \\
&= P(A_K = 1|H_K^*)\widehat{H}_K w_K(1, \widehat{H}_K)\left\{E[f(H_K)|H_K^*] - \widehat{H}_K'\beta_K^* + \widehat{H}_K'\left[\psi_K - \psi_K^*\right]\right\} = \Omega_K(H_K^*).
\end{aligned}
$$

By assumption, we have that $E[\Omega_K(H_K^*)] = 0$. Now, considering the first set of equations we get

$$
\begin{aligned}
&E[U_K^{(1)}(\beta_K^*, \psi_K^*)|H_K^*] \\
&= E\left[\widehat{H}_K w_K(A_K, \widehat{H}_K)\left\{E[f(H_K)|H_K^*] - \widehat{H}_K'\beta_K^* + A_K\widehat{H}_K'\left[\psi_K - \psi_K^*\right]\right\}\Big| H_K^*\right] \\
&= P(A_K = 1|H_K^*)\widehat{H}_K w_K(1, \widehat{H}_K)\left\{E[f(H_K)|H_K^*] - \widehat{H}_K'\beta_K^* + \widehat{H}_K'\left[\psi_K - \psi_K^*\right]\right\} \\
&\quad + P(A_K = 0|H_K^*)\widehat{H}_K w_K(0, \widehat{H}_K)\left\{E[f(H_K)|H_K^*] - \widehat{H}_K'\beta_K^*\right\} \\
&= \Omega_K(H_K^*) + P(A_K = 0|H_K^*)\widehat{H}_K w_K(0, \widehat{H}_K)\left\{E[f(H_K)|H_K^*] - \widehat{H}_K'\beta_K^*\right\}.
\end{aligned}
$$

Now, since $E[U_K^{(1)}(\beta_K^*, \psi_K^*)] = 0$ we get that,

$$
\begin{aligned}
0 &= E[\Omega_K(H_K^*)] + E\left[P(A_K = 0|H_K^*)\widehat{H}_K w_K(0, \widehat{H}_K)\left\{E[f(H_K)|H_K^*] - \widehat{H}_K'\beta_K^*\right\}\right] \\
&= E\left[P(A_K = 0|H_K^*)\widehat{H}_K w_K(0, \widehat{H}_K)\left\{E[f(H_K)|H_K^*] - \widehat{H}_K'\beta_K^*\right\}\right] \\
&= E\left[P(A_K = 1|H_K^*)\widehat{H}_K w_K(1, \widehat{H}_K)\left\{E[f(H_K)|H_K^*] - \widehat{H}_K'\beta_K^*\right\}\right], \qquad \text{(B.3.2)}
\end{aligned}
$$

where the last equality follows from (A2). Consider that

$$
\begin{aligned}
\Omega_K(H_K^*) = P(A_K = 1|H_K^*)\widehat{H}_K w_K(1, \widehat{H}_K)&\left\{E[f(H_K)|H_K^*] - \widehat{H}_K'\beta_K^*\right\} \\
&+ P(A_K = 1|H_K^*)\widehat{H}_K w_K(1, \widehat{H}_K)\widehat{H}_K'\left[\psi_K - \psi_K^*\right].
\end{aligned}
$$

The first line of this expression has zero expectation owing to Equation(B.3.2), which means that

$$
E\left[P(A_K = 1|H_K^*)\widehat{H}_K w_K(1, \widehat{H}_K)\widehat{H}_K'\right]\left[\psi_K - \psi_K^*\right] = 0,
$$

which (under the stated regularity conditions) gives $\psi_K^* = \psi_K$, as required. □

Proof of Theorem 6.4.1. In order to be valid for effect estimation, we view the outcome Y as equal to a baseline, treatment-free component plus the "blips" received for each treatment, across all stages. That is, we write

$$E[Y|H_K, A_K] = f_1(X_1) + \sum_{j=1}^{K} A_j H_j' \psi_j = f_K(H_K) + A_K H_K' \psi_K,$$

supposing linear blip terms for each stage. Note that in this formation, $f_K(H_K)$ is such that

$$E[f_K(H_K)|H_{K-1}, A_{K-1}] = f_{K-1}(H_{K-1}) + A_{K-1} H_{K-1}' \psi_{K-1}.$$

This continues though all other stages, $j = 1, \ldots, K-2$. Because of this, assuming the approximately non-differential error mechanism that has been discussed, we would also be able to write that,

$$E[f_K(H_K)|H_{K-1}^*, A_{K-1}] = E[f_{K-1}(H_{K-1})|H_{K-1}^*] + A_{K-1} \widehat{H}_{K-1}' \psi_{K-1}.$$

Then, so long as $E[V_{K-1}|H_{K-1}^*, A_{K-1}] = E[f_K(H_K)|H_{K-1}^*, A_{K-1}]$, the pseudo outcome can be validly used for effect estimation. This is equivalent to the claim that

$$E\left[A_K H_K' \psi_K - A_K \widehat{H}_K' \psi_K \middle| H_{K-1}^*, A_{K-1} \right] = 0.$$

This is easy to show through the law of iterated expectation, conditioning further on $\{H_K^*, A_K\}$. □

Proof. Proof of Theorem 6.4.2 Suppose that, as the theorem statement implies,

$$E[I(H_K' \psi_K > 0) H_K' \psi_K | H_K^*],$$

is known. In order to estimate the optimal DTR, we frame the observed outcome as the optimal outcome, plus the regrets from all stages. That is, we write

$$E[Y|H_K, A_K] = Y^{\text{opt}} + \sum_{j=1}^{K} (A_j^{\text{opt}} - A_j) H_j' \psi_j = Y^{\text{opt}}(H_K) + (A_j^{\text{opt}} - A_j) H_K' \psi_j,$$

under the assumption of linear blip terms at each stage. Moreover, we can write down that

$$E[Y^{\text{opt}}(H_K)|H_{K-1}, A_{K-1}] = Y^{\text{opt}}(H_{K-1}) + (A_{K-1}^{\text{opt}} - A_{K-1}) H_{K-1}' \psi_{K-1},$$

which is the necessary requirement to apply Theorem 6.3.1. To demonstrate that V_{K-1} is a valid pseudo outcome, notice that

$$\begin{aligned}
V_{K-1} &= Y - \left(E[I(H_K'\psi_K > 0)H_K'\psi_K|H_K^*] - A_K\widehat{H}_K'\psi_K\right) \\
&= Y^{\text{opt}}(H_K) + (A_K^{\text{opt}} - A_K)H_K'\psi_K \\
&\quad - \left(E[I(H_K'\psi_K > 0)H_K'\psi_K|H_K^*] - A_K\widehat{H}_K'\psi_K\right) \\
E[V_{K-1}|H_K^*, A_K^*] &= E[Y^{\text{opt}}(H_{K-1})|H_K^*, A_K^*] \\
&\quad - E\left[\left(A_K^{\text{opt}} - I(H_K'\psi_K > 0)\right)H_K'\psi_K\middle|H_K^*, A_K\right] \\
&= E[Y^{\text{opt}}(H_{K-1})|H_K^*, A_K^*] \\
E[V_{K-1}|H_{K-1}^*, A_{K-1}^*] &= E[Y^{\text{opt}}(H_{K-1})|H_{K-1}^*, A_{K-1}^*],
\end{aligned}$$

as required. □

Proof of Theorem 6.4.3. Consider

$$\begin{aligned}
E[A_K^{\text{opt}}H_K'\psi_K|H_K^*, A_K] &= E[I(H_K'\psi_K > 0)H_K'\psi_K|H_K^*] \\
&= E[I(Z_K + C_K > 0)(Z_K + C_K)|H_K^*] \\
&= P(Z_K > -C_K|H_K^*)\left\{C_K + E[Z_K|H_K^{*,\text{EP}}, Z_K > -C_K]\right\}.
\end{aligned}$$

Here we have defined $Z_K = H_K^{\text{EP}'}\psi_K^{\text{EP}}$ and $C_K = H_K^{\text{EF},T}\psi_K^{\text{EF}}$ to be the error-prone and error-free components of $H_K'\psi_K$. The assumptions on the error-prone covariates will allow us to conclude that $Z_K|H_K^{*,\text{EP}}$ is normally distributed, and correspondingly, the conditional expectation of Z_K is given by the mean of a truncated normal variable. This follows directly from the fact that both Z_K and $H_K^{*,\text{EP}}$ are joint combinations of $\begin{pmatrix}H_K^{\text{EP}'} & U'\end{pmatrix}'$, and so standard results of the multivariate normal distribution give the joint, and conditional distributions as being normal as well.

From the distributional assumption, U is normally distributed with mean 0 and variance Σ. Note that here the dimension of Σ_X and Σ will be identical. Then, we get that $Z_K|H_K^{*,\text{EP}}$ is distributed as,

$$N\left(\psi_K^{\text{EF}'}\mu_X + \psi_K^{\text{EF}'}\Sigma\left(\Sigma + \Sigma_X\right)^{-1}(H_K^{*,\text{EP}} - \mu_X), \psi_K^{\text{EF}'}\Sigma\psi_K^{\text{EF}} - \psi_K^{\text{EF}'}\Sigma(\Sigma + \Sigma_X)^{-1}\Sigma\psi_K^{\text{EF}}\right).$$

We denote these to be $\dot{\mu}_K$ and $\dot{\sigma}_K^2$. Then, using results from truncated normal distributions

we find that

$$E\left[Z_K|H_K^{*,\mathrm{EP}}, Z_K > -C_K\right] = \dot{\mu}_K + \dot{\sigma}_K \frac{\varphi\left(-\frac{\dot{\mu}_K + C_K}{\dot{\sigma}_K}\right)}{1 - \Phi\left(-\frac{\dot{\mu}_K + C_K}{\dot{\sigma}_K}\right)}.$$

From this argument we can also write that

$$P(Z_K > -C_K|H_K^*) = 1 - P(Z_K \leq -C_K|H_K^*) = 1 - \Phi\left(-\frac{C_K + \dot{\mu}_K}{\dot{\sigma}_K}\right).$$

Correspondingly, we find that

$$E[A_K^{\mathrm{opt}} H_K' \psi_K|H_K^*, A_K] = C_K \left\{1 - \Phi\left(-\frac{C_K + \dot{\mu}_K}{\dot{\sigma}_K}\right)\right\} + \dot{\mu}_K + \dot{\sigma}_K \varphi\left(-\frac{\dot{\mu}_K + C_K}{\dot{\sigma}_K}\right).$$

□

B.4 Chatper 7

Proof for Theorem 7.5.1. First, we show that

$$E[\widetilde{V}_{j+1}|H_j^*, A_j^*] = \nu_j^*(H_j^*) + \pi_j^*(H_j^*, A_j^*) C_j^*(H_j^*).$$

Then we show that $E[U_j^*(\psi_j)] = 0$. We begin using induction. First, for $j = K + 1$, we have $\widetilde{V}_j = Y$ and so

$$\begin{aligned}
&E[Y|H_K^*, A_K^*] \\
&= E\left\{E\left[Y|H_K, A_K, \overline{A}_K^*\right]\middle| H_K^*, A_K^*\right\} \\
&= E\left\{E\left[Y|H_K, A_K\right]| H_K^*, A_K^*\right\} && \text{I.A. (1)} \\
&= E\left\{Q_K(H_K, A_K)| H_K^*, A_K^*\right\} \\
&= E\left\{\nu_K(H_K) + A_K C_K(H_K; \psi_K)| H_K^*, A_K^*\right\} \\
&= E\left\{\nu_K(H_K)| H_K^*, A_K^*\right\} + E\left\{A_K C_K(H_K; \psi_K)| H_K^*, A_K^*\right\} \\
&= \nu_K^*(H_K^*) + P(A_K = 1|H_K^*, A_K^*) E\left\{C_K(H_K; \psi_K)| A_K = 1, H_K^*, A_K^*\right\} \\
&= \nu_K^*(H_K^*) + \pi_K^*(H_K^*, A_K^*) C_K^*(H_K^*) && \text{I.A. (2).}
\end{aligned}$$

Next, suppose take the inductive hypothesis (I.H.) to be that this expression holds for $j = K, \ldots, k+2$, and consider

$$
\begin{aligned}
&E[\widetilde{V}_{k+1}|H_k^*, A_k^*] \\
&= E[\widetilde{V}_{k+2} + \left\{A_{k+1}^{\text{opt}} - \pi_{k+1}^*(H_{k+1}^*, A_{k+1}^*)\right\} C_{k+1}^*(H_{k+1}^*)|H_k^*, A_k^*] \\
&= E\left\{ E\left[\widetilde{V}_{k+2}|H_{k+1}^*, A_{k+1}^*\right] + \left\{A_{k+1}^{\text{opt}} - \pi_{k+1}^*(H_{k+1}^*, A_{k+1}^*)\right\} C_{k+1}^*(H_{k+1}^*)\Big| H_k^*, A_k^*\right\} \\
&= E\left\{ \nu_{k+1}^*(H_{k+1}^*) + \pi_{k+1}^*(H_{k+1}^*, A_{k+1}^*)C_{k+1}^*(H_{k+1}^*)\right. \\
&\qquad \left. + \left\{A_{k+1}^{\text{opt}} - \pi_{k+1}^*(H_{k+1}^*, A_{k+1}^*)\right\} C_{k+1}^*(H_{k+1}^*)\Big| H_k^*, A_k^*\right\} && \text{I.H} \\
&= E\left\{ \nu_{k+1}^*(H_{k+1}^*) + A_{k+1}^{\text{opt}} C_{k+1}^*(H_{k+1}^*)\Big| H_k^*, A_k^*\right\} \\
&= E\left\{ E[\nu_{k+1}(H_{k+1}) + A_{k+1}^{\text{opt}} C_{k+1}(H_{k+1})|H_{k+1}^*, A_{k+1}^*]\Big| H_k^*, A_k^*\right\} \\
&= E\left\{ \nu_{k+1}(H_{k+1}) + A_{k+1}^{\text{opt}} C_{k+1}(H_{k+1})\Big| H_k^*, A_k^*\right\} \\
&= E\left\{ V_{k+1}(H_{k+1})\right| H_k^*, A_k^*\} \\
&= E\left\{ E\left\{V_{k+1}(H_{k+1})|H_k, A_k, \overline{A}_k^*\right\}\Big| H_k^*, A_k^*\right\} \\
&= E\left\{ \nu_k(H_k) + A_k C_k(H_k)\right| H_k^*, A_k^*\} && \text{I.A. (1)} \\
&= \nu_k^*(H_k^*) + P(A_k = 1|H_k^*, A_k^*) E\left\{ C_k(H_k)\right| A_k = 1, H_k^*, A_k^*\} \\
&= \nu_k^*(H_k^*) + \pi_k^*(H_k^*, A_k^*) C_k(H_k^*). && \text{I.A. (2)}
\end{aligned}
$$

Note that in addition to the independence assumptions and the inductive hypothesis, we also used the fact that $C_k(\cdot)$ is correctly specified. In the event (as will be the case in practice) that we are using the estimated versions instead, all of these equalities hold almost surely (assuming that $\widehat{\psi}_j$ are almost surely consistent for ψ_j). With these expected pseudo-outcome (E.P.O) results established, we can show that $E[U_j^*(\psi_j)] = 0$. First, consider the expectation, conditional on $\{H_j^*, A_j^*\}$

$$
\begin{aligned}
&E\left[U_j^*(\psi_j)|H_j^*, A_j^*\right] \\
&= \sum_{i=1}^n \lambda_j^*(H_{i,j}^*)\left\{A_{i,j}^* - P(A_{i,j}^* = 1|H_{i,j}^*)\right\} \\
&\qquad \times \left\{E[\widetilde{V}_{i,j+1}|H_{i,j}^*, A_{i,j}^*] - \pi_j^*(H_{i,j}^*, A_{i,j}^*)C_j^*(H_{i,j}^*; \psi_j) + \theta_j^*(H_{i,j}^*)\right\} \\
&= \sum_{i=1}^n \lambda_j^*(H_{i,j}^*)\left\{A_{i,j}^* - P(A_{i,j}^* = 1|H_{i,j}^*)\right\}\left\{\nu_j^*(H_{i,j}^*) + \theta_j^*(H_{i,j}^*)\right\} && \text{E.P.O.}
\end{aligned}
$$

Taking this result, we can consider the expectation conditional on just H_j^*, which gives

$$\begin{aligned}
&E\left[U_j^*(\psi_j)|H_j^*\right] \\
&= \sum_{i=1}^{n} \lambda_j^*(H_{i,j}^*)\left\{E[A_{i,j}^*|H_{i,j}^*] - P(A_{i,j}^* = 1|H_{i,j}^*)\right\}\left\{\nu_j^*(H_{i,j}^*) + \theta_j^*(H_{i,j}^*)\right\} \qquad \text{I.A. (3)} \\
&= \sum_{i=1}^{n} \lambda_j^*(H_{i,j}^*)\left\{P(A_{i,j}^* = 1|H_{i,j}^*) - P(A_{i,j}^* = 1|H_{i,j}^*)\right\}\left\{\nu_j^*(H_{i,j}^*) + \theta_j^*(H_{i,j}^*)\right\} \\
&= 0.
\end{aligned}$$

Note that this will hold so long as the residual term

$$E[\widetilde{V}_{i,j+1}|H_{i,j}^*, A_{i,j}^*] - \pi_j^*(H_{i,j}^*, A_{i,j}^*)C_j^*(H_{i,j}^*; \psi_j) + \theta_j^*(H_{i,j}^*),$$

is independent of $H_{i,j}^*$. This is true under the assumptions laid out only at the true ψ_j, in general. As a result, $U_j^*(\psi_j)$ form unbiased estimating equations which are uniquely solved at the true ψ_j, and as a result produce consistent estimators for ψ_j. □

Proof of Theorem 7.7.1. Under non-exceptional laws, and the standard regularity conditions, then Theorem 7.5.1 demonstrates that the U^* is an unbiased estimating equation. Supposing that the other nuisance parameters are estimated via M-estimation techniques, then a simple invocation of two-step M-estimation theory provides the necessary results. □

Appendix C

Non-regularity in DTRs

We mentioned that the asymptotic theory for DTR estimation is subject to non-regularity. We present an illustrative example of this here. Consider a simple example of a non-regular estimator (largely taken from Tsiatis, Davidian, Holloway, and Laber [90], though there is a mistake in the published argument). Take H_1 to be scalar valued, with $\mathcal{A}_1 = \{0, 1\}$. Assume that $Q_1(h_1, a_1; \beta_1)$ is correctly specified as $\beta_{10} + \beta_{11}h_1 + \beta_{12}a_1$. This gives the optimal treatment rule, $d_1^{\text{opt}} = I(\beta_{12} > 0)$, and leads to $\max_{a_1} Q_1(h_1, a_1; \beta_1) = \beta_{10} + \beta_{11}h_1 + \beta_{12}I(\beta_{12} > 0)$. We can derive the value of this to be given by $\mathcal{V}(d^{\text{opt}}) = \beta_{10} + \beta_{11}E[H_1] + \beta_{12}I(\beta_{12} > 0)$. The estimated value, based on the sample, will be given by $\widehat{\mathcal{V}} = \widehat{\beta}_{10} + \widehat{\beta}_{11}\overline{H}_1 + \widehat{\beta}_{12}I(\widehat{\beta}_{12} > 0)$. Then,

$$
\begin{aligned}
\widehat{\mathcal{V}} - \mathcal{V} &= \widehat{\beta}_{10} + \widehat{\beta}_{11}\overline{H}_1 + \widehat{\beta}_{12}I(\widehat{\beta}_{12} > 0) - \{\beta_{10} + \beta_{11}E[H_1] + \beta_{12}I(\beta_{12} > 0)\} \\
&= \widehat{\beta}_{10} + \widehat{\beta}_{11}\overline{H}_1 + \widehat{\beta}_{12}I(\widehat{\beta}_{12} > 0) - \{\beta_{10} + \beta_{11}E[H_1] + \beta_{12}I(\beta_{12} > 0)\} \\
&+ \widehat{\beta}_{11}\left(E[H_1] - E[H_1]\right) \\
&= \left(\widehat{\beta}_{10} - \beta_{10}\right) + \left(\widehat{\beta}_{11} - \beta_{11}\right)E[H_1] + \left(\widehat{\beta}_{12}I(\widehat{\beta}_{12} > 0) - \beta_{12}I(\beta_{12} > 0)\right) \\
&+ \widehat{\beta}_{11}\left(\overline{H}_1 - E[H_1]\right).
\end{aligned}
$$

Adding and subtracting $\beta_{11}\left(\overline{H}_1 - E[H_1]\right)$, and multiplying by $\sqrt{n}$, gives

$$\begin{aligned}\sqrt{n}\left(\widehat{\mathcal{V}} - \mathcal{V}\right) &= \sqrt{n}\left(\widehat{\beta}_{10} - \beta_{10}\right) + \sqrt{n}\left(\widehat{\beta}_{11} - \beta_{11}\right)E[H_1] \\ &+ \sqrt{n}\left(\widehat{\beta}_{12}I(\widehat{\beta}_{12} > 0) - \beta_{12}I(\beta_{12} > 0)\right) \\ &+ \sqrt{n}\left(\widehat{\beta}_{11} - \beta_{11}\right)\left(\overline{H}_1 - E[H_1]\right) \\ &+ \sqrt{n}\beta_{11}\left(\overline{H}_1 - E[H_1]\right).\end{aligned}$$

Standard M-estimator theory gives us that

$$\sqrt{n}\begin{pmatrix}\widehat{\beta}_{10} - \beta_{10} \\ \widehat{\beta}_{11} - \beta_{11} \\ \widehat{\beta}_{12} - \beta_{12} \\ \overline{H}_1 - E[H_1]\end{pmatrix} \xrightarrow{d} \begin{pmatrix}Z_1 \\ Z_2 \\ Z_3 \\ Z_4\end{pmatrix} N(0, \Sigma).$$

The fact that $\overline{H} \to E[H_1]$ means that the term on the third line converges in probability to 0, the top line will converge in distribution to $Z_1 + E[H_1]Z_2$, and the final component to $\beta_{11}Z_4$. This leaves only $\sqrt{n}\left(\widehat{\beta}_{12}I(\widehat{\beta}_{12} > 0) - \beta_{12}I(\beta_{12} > 0)\right)$. $g(u)$ is not differentiable at $u = 0$, and so the Delta Method can only be applied if $\beta_{12} \neq 0$. Making this assumption, this term will converge in distribution to $Z_{12}I(\beta_{13} > 0)$, so that

$$\sqrt{n}\left(\widehat{\mathcal{V}} - \mathcal{V}\right) \xrightarrow{d} Z_1 + E[H_1]Z_2 + I(\beta_{12} > 0)Z_3 + \beta_{11}Z_4.$$

If instead we have $\beta_{12} = 0$, then this term simplifies to $\sqrt{n}\widehat{\beta}_{12}I(\widehat{\beta}_{12} > 0)$, and since $\beta_{12} = 0$ we have $\sqrt{n}\widehat{\beta}_{12} \xrightarrow{d} Z_3$. Since the indicator $I(\widehat{\beta}_{12} > 0) = I(\sqrt{n}\widehat{\beta}_{12} > 0)$ we can apply the continuous mapping theorem, resulting in $\sqrt{n}\widehat{\beta}_{12}I(\widehat{\beta}_{12} > 0) \xrightarrow{d} Z_3I(Z_3 > 0)$. As a result,

$$\sqrt{n}\left(\widehat{\mathcal{V}} - \mathcal{V}\right) \xrightarrow{d} \begin{cases} Z_1 + E[H_1]Z_2 + I(\beta_{12} > 0)Z_3 + \beta_{11}Z_4 & \beta_{12} \neq 0 \\ Z_1 + E[H_1]Z_2 + I(Z_3)Z_3 + \beta_{11}Z_4 & \text{otherwise} \end{cases},$$

where in the first case the limiting distribution is normal and in the second case it is not. While this is a specific realization of the problem of non-regularity in DTRs, these issues ultimately stem from the violation of regularity conditions we previously discussed. In this case, $\widehat{\mathcal{V}}$ is subject to standard asymptotic theory whenever $\beta_{11} \neq 0$, however, whenever there is no treatment effect, non-regular theory will be necessary.

Appendix D

Additional Simulation Results

When investigating the impact of measurement error in dynamic treatment regimes, in Chapter 6, we simulated numerous additional scenarios to investigate the impact of various factors on our proposed corrections. The results of these simulations are provided here. To investigate the procedure in the multistage scenario, we considered a variety of related settings formed by varying different aspects of the model. We take $X_1 \sim N(0,1)$, with $X_{11}^* \sim g_1(X_1)$ and $X_{12}^* \sim g_2(X_1)$, for error models g_1, g_2. The $P(A_1 = 1|X_1^* = x_1^*) = h_1(x_1^*; \alpha_{10}, \alpha_{11})$, with a treatment model h_1 and parameters α_{10}, α_{11}. We then take $X_2 \sim N(A_1, 1)$, with $X_{21}^* = g_1(X_2)$ and $X_{22}^* = g_2(X_2)$, and $P(A_2 = 1|X_2^* = x_2^*) = h_2(x_2^*; \alpha_{20}, \alpha_{21})$. The outcome is then given by $Y = f(X_1) + (A_1^{\text{opt}} - A_1)(1 + \psi_{11}X_1) + (A_2^{\text{opt}} - A_2)(1 + \psi_{21}X_2) + \epsilon$, with $\epsilon \sim N(0,1)$, where $f(X_1)$ is the treatment-free model. The five considered scenarios depend on the alteration of the above parameters.

1. Considers 10 combinations of $(\alpha_{10}, \alpha_{20})$, values taken from $\{-2, -1, 0, 1, 2\}$, holding the treatment-free model as linear, both treatment models as linear, the error models as classical additive with $N(0, 0.25)$ distribution, $\psi_{11} = \psi_{21} = 1$.

2. Considers 10 combinations of (ψ_{11}, ψ_{21}), values taken from $\{-1, -0.1, 0, 0.1, 1\}$, holding $\alpha_{10} = \alpha_{20} = 0$, the treatment-free model as linear, both treatment models as linear, the error models as classical additive with $N(0, 0.25)$ distribution.

3. Considers 5 scenarios for various forms of the treatment-free model, taking $f(X_1) = X_1$ (linear), $f(X_1) = X_1 + X_1^2$ (quadratic), $f(X_1) = X_1 + X_2^2 - X_1^3$ (cubic), $f(X_1) = \exp(X_1) - X_1^3$ (exponential), or $\exp(X_1)I(X_1 >= -0.5)$ (complex). We hold both treatment models to be linear, $\alpha_{10} = \alpha_{20} = 0$, $\psi_{11} = \psi_{21} = 1$, and the error models as classical additive with $N(0, 0.25)$ distribution.

4. Considers 10 scenarios where the treatment models are taken to be one of $h_j(x_j^*) = \alpha_{j0} + \alpha_{j1}x_j^*$ (linear), $h_j(x_j^*) = \alpha_{j0} + \alpha_{j1}x_j^* + (x_j^*)^2$ (quadratic), $h_j(x_j^*) = \alpha_{j0} + \alpha_{j1}x_j^* + \exp(x_j^*)$ (exponential), and $h_j(x_j^*) = \alpha_{j0} + \alpha_{j1}x_j^* + (x_j^*)^2 + \exp(x_j^*)$ (mixed). We hold the treatment-free model to be linear, $\alpha_{10} = \alpha_{20} = 0$, $\psi_{11} = \psi_{21} = 1$, and the error models as classical additive with $N(0, 0.25)$ distribution.

5. Considers 10 scenarios for various error models, taking $g_j(X_l) = X_l + N(0, 0.25)$ (normal), $g_j(X_l) = X_l + t_{10}$ (approximately normal), $g_j(X_l) = X_l \cdot \text{Gamma}(1, 1)$ (gamma), or $g_j(X_l) = X_l \cdot \text{Unif}(0.5, 1.5)$ (uniform). We hold the treatment-free model to be linear, both treatment models to be linear, $\alpha_{10} = \alpha_{20} = 0$, and $\psi_{11} = \psi_{21} = 1$.

All analyses were conducted where (X_1^*, X_2^*) is taken to be

$$(X_{11}^*, X_{21}^*), (\overline{X_1^*}, \overline{X_2^*}), (X_{11}^*, \overline{X_2^*}), (\overline{X_1^*}, X_{21}^*)$$

(that is treatment depends on either the first naive proxy, or on the mean of the two proxies). We take $n = 10,000$ and repeat each scenario 1000 times. The results for a corrected analysis and a naive analysis are included in tables D.1-D.5.

Table D.1: Median parameter estimates investigating the impact of treatment probabilities in a multistage DTR, by varying $(\alpha_{10}, \alpha_{20})$ as indicated. Blip parameter estimates are compared for $n = 10,000$ individuals, using the corrected method compared to a naive analysis. The top set of rows of the table use the first error-prone proxy at both stages, the second set of rows use the mean of proxies at both stages, the third set of rows use the mean at the first stage and the first error-prone proxy at the second, and the final set of rows use the first error-prone proxy at the first stage and the mean at the second. Bold values indicate parameters for which the 95% percentile-based interval across the 1000 simulation replicates did **not** cover the true parameter value.

	Regression Calibration				Naive			
$(\alpha_{10}, \alpha_{20})$	A_1	A_1X_1	A_2	A_2X_2	A_1	A_1X_1	A_2	A_2X_2
(-2, -2)	1.0101	0.9974	1.0035	0.9963	1.0098	**0.8874**	1.0163	**0.902**
(-1, -1)	1.0103	0.9983	1.0038	0.9961	1.0102	**0.8879**	1.0304	**0.9064**
(0, 0)	1.0106	0.9961	1.0003	0.9995	1.0108	**0.8859**	1.045	**0.9101**
(1, 1)	1.0086	0.996	0.998	0.9994	1.0086	**0.8857**	**1.0622**	**0.909**
(2, 2)	1.0082	0.9983	0.9975	1.0027	1.0083	**0.8881**	1.0789	**0.9077**
(-2, 0)	1.01	0.9973	1.0009	0.9987	1.01	**0.8872**	1.0157	**0.8993**
(-1, 1)	1.0109	0.9977	0.999	0.9986	1.0109	**0.8868**	1.0267	**0.9044**

	Regression Calibration				Naive			
$(\alpha_{10}, \alpha_{20})$	A_1	A_1X_1	A_2	A_2X_2	A_1	A_1X_1	A_2	A_2X_2
(0, 2)	1.011	0.9963	0.9963	0.9987	1.0105	**0.8859**	1.0413	**0.9079**
(1, -2)	1.0087	0.9971	1.0028	0.9993	1.0084	**0.886**	1.0704	**0.904**
(2, -1)	1.0084	0.9997	1.0036	0.9975	1.0079	**0.8886**	**1.0891**	**0.8982**
(-2, -2)	1.008	0.9992	1.0034	0.9965	1.0087	**0.8887**	1.0165	**0.902**
(-1, -1)	1.0112	0.9988	1.0048	0.998	1.0107	**0.8877**	1.031	**0.9068**
(0, 0)	1.0111	0.997	1.0013	0.9967	1.0111	**0.8872**	1.0454	**0.9074**
(1, 1)	1.0093	0.9963	0.9975	0.9985	1.009	**0.8865**	**1.0621**	**0.9083**
(2, 2)	1.0101	0.9993	0.9977	1.0007	1.0104	**0.8884**	1.0795	**0.9054**
(-2, 0)	1.0108	0.9994	1.0011	0.9985	1.0108	**0.8884**	1.0159	**0.8991**
(-1, 1)	1.011	0.9982	0.9979	0.9988	1.0113	**0.8875**	1.0257	**0.9039**
(0, 2)	1.0119	0.9972	0.9956	0.9972	1.0118	**0.8865**	1.041	**0.9073**
(1, -2)	1.0091	0.9973	1.0043	0.9976	1.009	**0.8866**	1.072	**0.903**
(2, -1)	1.01	1.0002	1.0014	0.998	1.0091	**0.8894**	**1.0871**	**0.8974**
(-2, -2)	1.0083	1.0001	1.0038	0.9975	1.0082	**0.8895**	1.0163	**0.9018**
(-1, -1)	1.0107	0.999	1.0039	0.9971	1.0106	**0.8882**	1.0305	**0.9064**
(0, 0)	1.0116	0.9968	1.0005	0.9982	1.0118	**0.8866**	1.0453	**0.9097**
(1, 1)	1.0096	0.9955	0.9981	0.9995	1.0095	**0.8862**	**1.0619**	**0.909**
(2, 2)	1.0101	0.9994	0.9968	1.0011	1.0098	**0.8889**	1.0802	**0.9066**
(-2, 0)	1.0107	0.9989	1.0008	0.9991	1.0113	**0.8886**	1.0156	**0.8998**
(-1, 1)	1.0113	0.9983	0.9989	0.9995	1.011	**0.8874**	1.0262	**0.9055**
(0, 2)	1.0116	0.9969	0.9967	0.9976	1.0115	**0.8869**	1.042	**0.9081**
(1, -2)	1.0093	0.9975	1.0043	0.999	1.0093	**0.8869**	1.0721	**0.9041**
(2, -1)	1.0098	0.9996	1.0024	0.9975	1.0097	**0.8891**	**1.0887**	**0.8974**
(-2, -2)	1.0096	0.9977	1.0034	0.9979	1.0095	**0.8875**	1.0162	**0.9027**
(-1, -1)	1.0094	0.9978	1.0038	0.9976	1.0097	**0.8872**	1.0304	**0.9067**
(0, 0)	1.0106	0.9963	1.0013	0.9974	1.0104	**0.8859**	1.0459	**0.9085**
(1, 1)	1.0085	0.9965	0.9982	1.0001	1.0084	**0.8857**	**1.0621**	**0.9093**
(2, 2)	1.0079	0.9983	0.9961	0.9983	1.0078	**0.8879**	1.0786	**0.9043**
(-2, 0)	1.0099	0.9983	1.0007	0.9977	1.01	**0.8872**	1.0155	**0.8988**
(-1, 1)	1.0112	0.9976	0.9983	0.9981	1.011	**0.8866**	1.0262	**0.9043**
(0, 2)	1.0108	0.9962	0.9959	0.9957	1.0108	**0.8863**	1.0411	**0.906**
(1, -2)	1.0078	0.9973	1.0034	0.998	1.0074	**0.8864**	1.071	**0.9031**
(2, -1)	1.0082	0.9996	1.0036	0.9969	1.0079	**0.8887**	**1.0881**	**0.8972**

Table D.2: Median parameter estimates investigating the impact of treatment thresholds in a multistage DTR, by varying (ψ_{11}, ψ_{21}) as indicated. Blip parameter estimates are compared for $n = 10,000$ individuals, using the corrected method compared to a naive analysis. The top set of rows of the table use the first error-prone proxy at both stages, the second set of rows use the mean of proxies at both stages, the third set of rows use the mean at the first stage and the first error-prone proxy at the second, and the final set of rows use the first error-prone proxy at the first stage and the mean at the second. Bold values indicate parameters for which the 95% percentile-based interval across the 1000 simulation replicates did **not** cover the true parameter value. Scenarios take $(\psi_{11}, \psi_{20}, \psi_{21})$ to be 1: (-1, 1, -1), 2: (-.1, 1, -.1), 3: (0, 1, 0), 4: (.1, 1, .1), 5: (1, 1, 1), 6: (-1, 1, 0), 7: (-.1, 1, .1), 8: (0, 1, 1), 9: (.1, 1, -1), and 10: (1, 1, -.1).

	Regression Calibration				Naive			
	A_1	A_1X_1	A_2	A_2X_2	A_1	A_1X_1	A_2	A_2X_2
1	0.9902	-1.001	1.0011	-1.002	0.9907	**-0.8904**	0.956	**-0.9114**
2	0.9999	-0.1017	1.0017	-0.1006	0.9998	-0.0904	0.9971	-0.0914
3	1	-0.0019	1.0014	-8e-04	1	-0.0017	1.0016	-9e-04
4	0.9998	0.0979	1.0016	0.099	0.9998	0.0872	1.006	0.0904
5	1.0106	0.9961	1.0003	0.9995	1.0108	**0.8859**	1.045	**0.9101**
6	1.0006	-1.0014	1.0017	-1e-04	1.0005	**-0.89**	1.0017	2e-04
7	0.9998	-0.1018	1.0017	0.0994	0.9998	-0.0905	1.0059	0.0906
8	1.0106	-0.0029	1.0017	0.9986	1.0105	-0.0026	1.0461	**0.9094**
9	0.9918	0.0988	1.0019	-1.0028	0.9918	0.0877	0.9563	**-0.9121**
10	0.9998	0.9967	1.0002	-0.1017	1.0004	**0.8867**	0.9958	-0.0922
1	0.993	-1.0006	0.9997	-1.0023	0.9934	**-0.8895**	0.9548	**-0.911**
2	1.0001	-0.1017	0.9993	-0.1018	1.0001	-0.0903	0.9951	-0.0924
3	1.0006	-0.0017	0.9995	-0.0017	1.0006	-0.0015	0.9996	-0.0018
4	1.0007	0.0985	0.9999	0.0976	1.0007	0.0875	1.0046	0.0889
5	1.0111	0.997	1.0013	0.9967	1.0111	**0.8872**	1.0454	**0.9074**
6	1.0025	-1.0015	0.9996	-0.0015	1.0025	**-0.8903**	0.9994	-0.0012
7	1.0006	-0.1016	0.9993	0.0981	1.0006	-0.0903	1.004	0.0895
8	1.0117	-0.0015	1.0016	0.9968	1.0117	-0.0013	1.0459	**0.9076**
9	0.9924	0.0992	1.0005	-1.0034	0.9925	0.0881	0.9549	**-0.9123**
10	0.9997	0.9969	1.0006	-0.1023	1.0001	**0.8871**	0.9957	-0.0933
1	0.9926	-1.0004	1.0011	-1.0041	0.9927	**-0.8893**	0.9556	**-0.913**
2	1.0015	-0.1013	1	-0.1008	1.0015	-0.09	0.9959	-0.092
3	1.0015	-0.0017	1.0001	-7e-04	1.0015	-0.0015	1	-8e-04

	Regression Calibration				Naive			
	A_1	A_1X_1	A_2	A_2X_2	A_1	A_1X_1	A_2	A_2X_2
4	1.0018	0.0984	1.0002	0.0991	1.0017	0.0875	1.0046	0.0902
5	1.0116	0.9968	1.0005	0.9982	1.0118	**0.8866**	1.0453	**0.9097**
6	1.0023	-1.0016	1.0008	-0.0011	1.0022	**-0.8901**	1.0008	-0.0015
7	1.0016	-0.1016	1.0006	0.0994	1.0016	-0.0903	1.0048	0.09
8	1.0121	-0.0014	1.0004	0.9982	1.0121	-0.0012	1.045	**0.9097**
9	0.9926	0.0996	0.9999	-1.0028	0.9927	0.0885	0.9549	**-0.9121**
10	1.001	0.9971	0.9998	-0.1007	1.0009	**0.8872**	0.9952	-0.0919
1	0.9893	-1.0008	1.0004	-1.0018	0.9893	**-0.8897**	0.9554	**-0.9112**
2	0.9992	-0.1017	1.0013	-0.1014	0.9992	-0.0906	0.9964	-0.0928
3	0.9994	-0.0019	1.0013	-0.0016	0.9993	-0.0017	1.001	-0.0018
4	0.9997	0.0979	1.0012	0.0982	0.9996	0.0873	1.0058	0.0893
5	1.0106	0.9963	1.0013	0.9974	1.0104	**0.8859**	1.0459	**0.9085**
6	1.0008	-1.0013	0.9997	-6e-04	1.0009	**-0.8901**	0.9996	-8e-04
7	0.9997	-0.1014	1.001	0.0987	0.9996	-0.0903	1.0056	0.0894
8	1.0102	-0.0023	1.0011	0.9986	1.0102	-0.002	1.0456	**0.9091**
9	0.9914	0.0986	1.0013	-1.0021	0.9914	0.0876	0.9563	**-0.9119**
10	0.9994	0.9966	1.0005	-0.1016	0.9995	**0.8867**	0.9958	-0.0925

Table D.3: Median parameter estimates investigating the impact of treatment probabilities in a multistage DTR, by varying the true treatment-free model as indicated. Linear treatment-free models are used in all settings. Blip parameter estimates are compared for $n = 10,000$ individuals, using the corrected method compared to a naive analysis. The top set of rows of the table use the first error-prone proxy at both stages, the second set of rows use the mean of proxies at both stages, the third set of rows use the mean at the first stage and the first error-prone proxy at the second, and the final set of rows use the first error-prone proxy at the first stage and the mean at the second. Bold values indicate parameters for which the 95% percentile-based interval across the 1000 simulation replicates did **not** cover the true parameter value.

	Regression Calibration				Naive			
TF Model	A_1	A_1X_1	A_2	A_2X_2	A_1	A_1X_1	A_2	A_2X_2
Linear	1.0106	0.9961	1.0003	0.9995	1.0108	**0.8859**	1.045	**0.9101**
Quadratic	1.0105	1.0043	1.0012	0.9982	1.0104	**0.893**	1.0457	**0.9098**
Cubic	1.0019	1.0108	1.0041	0.9993	1.0021	0.8986	1.0484	0.9099

	Regression Calibration				Naive			
TF Model	A_1	A_1X_1	A_2	A_2X_2	A_1	A_1X_1	A_2	A_2X_2
Exponential	1.0042	1.0105	1.0035	0.9986	1.0041	0.8976	1.0476	0.9104
Complex	1.0125	1.0038	0.9994	1.0007	1.0127	0.8921	1.044	**0.9116**
Linear	1.0111	0.997	1.0013	0.9967	1.0111	**0.8872**	1.0454	**0.9074**
Quadratic	1.0106	1.0017	1.0033	0.9959	1.0107	**0.8901**	1.0483	**0.9078**
Cubic	1.0069	1.0043	1.0072	1.0012	1.0068	0.8938	1.0525	0.9129
Exponential	1.0089	1.0066	1.0054	1.0015	1.0091	0.8939	1.0503	0.9122
Complex	1.0113	1.0014	1.0035	0.9969	1.0116	**0.8903**	1.0467	**0.9082**
Linear	1.0116	0.9968	1.0005	0.9982	1.0118	**0.8866**	1.0453	**0.9097**
Quadratic	1.0116	1.001	1.0005	0.9985	1.0115	**0.8897**	1.044	**0.9093**
Cubic	1.0075	1.0034	1.0038	0.9967	1.0075	0.8937	1.0484	0.9095
Exponential	1.0091	1.0069	1.0033	0.9996	1.0087	0.8949	1.0477	0.9107
Complex	1.0119	1.0026	1	1.0012	1.0116	**0.8904**	1.0445	**0.9108**
Linear	1.0106	0.9963	1.0013	0.9974	1.0104	**0.8859**	1.0459	**0.9085**
Quadratic	1.01	1.004	1.003	0.997	1.0104	**0.8927**	1.0472	**0.9078**
Cubic	1.0017	1.0101	1.0044	0.9992	1.0021	0.8974	1.0504	0.9104
Exponential	1.0046	1.0097	1.0061	0.9996	1.0045	0.8975	1.0503	0.9104
Complex	1.0115	1.0039	1.0015	0.9985	1.0115	0.8926	1.0457	**0.9091**

Table D.4: Median parameter estimates investigating the impact of treatment probabilities in a multistage DTR, by varying the treatment models as indicated. Linear treatment models are used in all situations. Blip parameter estimates are compared for $n = 10,000$ individuals, using the corrected method compared to a naive analysis. The top set of rows of the table use the first error-prone proxy at both stages, the second set of rows use the mean of proxies at both stages, the third set of rows use the mean at the first stage and the first error-prone proxy at the second, and the final set of rows use the first error-prone proxy at the first stage and the mean at the second. Bold values indicate parameters for which the 95% percentile-based interval across the 1000 simulation replicates did **not** cover the true parameter value. The treatment models are specified to be linear (L), quadratic (Q), mixed (M), or exponential (E).

	Regression Calibration				Naive			
	A_1	A_1X_1	A_2	A_2X_2	A_1	A_1X_1	A_2	A_2X_2
L/L	1.0106	0.9961	1.0003	0.9995	1.0108	**0.8859**	1.045	**0.9101**
L/Q	0.9992	0.9992	**0.8902**	**1.1138**	0.9991	**0.8881**	**0.9407**	1.014

	Regression Calibration				Naive			
	A_1	A_1X_1	A_2	A_2X_2	A_1	A_1X_1	A_2	A_2X_2
L/M	0.9925	1	**0.9022**	**1.1187**	0.9923	**0.8893**	0.9529	1.0174
L/E	1.0126	0.9989	0.9836	1.016	1.0124	**0.8876**	1.0291	0.9257
Q/Q	**0.9047**	**1.1117**	**0.8929**	**1.1153**	**0.9044**	0.9885	0.9611	1.0159
Q/M	**0.8969**	**1.1119**	**0.9011**	**1.1188**	**0.897**	0.9886	0.9707	1.0199
Q/E	**0.918**	**1.1125**	0.9846	1.0189	**0.918**	0.9899	1.0478	0.9289
M/M	**0.9033**	**1.1052**	**0.9078**	1.1216	**0.9033**	0.9824	0.9946	1.0192
M/E	**0.9246**	**1.1096**	0.9853	1.0153	**0.9244**	0.9872	1.0652	0.9223
E/E	0.9979	1.013	0.9853	1.0131	0.9983	**0.9007**	1.0479	0.9244
L/L	1.0111	0.997	1.0013	0.9967	1.0111	**0.8872**	1.0454	**0.9074**
L/Q	1.0013	0.9987	**0.8847**	**1.1219**	1.0012	**0.8876**	**0.9353**	1.0215
L/M	0.9915	0.9979	**0.8915**	**1.1179**	0.9916	**0.8868**	0.9413	1.0179
L/E	1.0128	0.999	0.9818	1.0194	1.013	**0.8882**	1.0271	0.9275
Q/Q	**0.898**	**1.1245**	**0.8863**	**1.1239**	**0.8982**	0.9994	0.9547	1.0234
Q/M	**0.8889**	**1.1226**	**0.8943**	**1.1241**	**0.8887**	0.9981	0.9629	1.0257
Q/E	**0.9089**	**1.121**	0.9809	1.021	**0.9087**	0.9965	1.0429	0.932
M/M	**0.8941**	**1.1119**	**0.9053**	1.1241	**0.8938**	0.9878	0.9923	1.0228
M/E	**0.9142**	**1.1125**	0.9834	1.0242	**0.9141**	0.9884	1.0622	0.9313
E/E	0.9945	1.0162	0.9831	1.0186	0.9948	**0.9021**	1.0474	0.9294
L/L	1.0116	0.9968	1.0005	0.9982	1.0118	**0.8866**	1.0453	**0.9097**
L/Q	1.0012	0.9979	**0.8904**	**1.1137**	1.0008	**0.8876**	**0.9412**	1.0139
L/M	0.9922	0.9985	**0.8999**	**1.1146**	0.9923	**0.8881**	0.9495	1.0143
L/E	1.011	0.9982	0.9823	1.0161	1.0112	**0.8869**	1.0281	0.9263
Q/Q	**0.8978**	**1.1239**	**0.8921**	**1.1151**	**0.8982**	0.9981	0.9598	1.0149
Q/M	**0.8904**	**1.1216**	**0.9021**	**1.1175**	**0.8903**	0.9979	0.9703	1.0182
Q/E	**0.9086**	**1.1215**	0.9848	1.0112	**0.9083**	0.9972	1.0464	0.923
M/M	**0.8932**	**1.1145**	**0.9077**	1.1196	**0.8937**	0.9898	0.9943	1.0184
M/E	**0.9133**	**1.1123**	0.9854	1.0169	**0.9126**	0.9888	1.0637	0.9247
E/E	0.996	1.0178	0.9844	1.0135	0.9955	**0.905**	1.0486	0.9253
L/L	1.0106	0.9963	1.0013	0.9974	1.0104	**0.8859**	1.0459	**0.9085**
L/Q	1.0003	0.9973	**0.884**	**1.1229**	0.9997	**0.8868**	**0.9348**	1.0231
L/M	0.9908	0.998	**0.8919**	**1.1214**	0.9904	**0.8869**	0.9419	1.0213
L/E	1.0111	0.9987	0.9803	1.0186	1.0113	**0.888**	1.0264	0.9292
Q/Q	**0.9049**	**1.1118**	**0.8872**	**1.1245**	**0.9041**	0.9884	0.9563	1.0241
Q/M	**0.8951**	**1.1141**	**0.8978**	1.1272	**0.8956**	0.9901	0.9677	1.0279
Q/E	**0.9178**	**1.1119**	0.9811	1.0225	**0.9179**	0.9884	1.0437	0.9333

	Regression Calibration				Naive			
	A_1	A_1X_1	A_2	A_2X_2	A_1	A_1X_1	A_2	A_2X_2
M/M	**0.9008**	**1.1039**	**0.9066**	1.1307	**0.901**	0.9811	0.9946	1.0289
M/E	**0.9248**	**1.1063**	0.9843	1.023	**0.9249**	0.9835	1.0644	0.9311
E/E	0.9971	1.0136	0.984	1.0232	0.997	**0.9008**	1.0482	0.9345

Table D.5: Median parameter estimates investigating the impact of treatment probabilities in a multistage DTR, by varying the error-models as indicated. Blip parameter estimates are compared for $n = 10,000$ individuals, using the corrected method compared to a naive analysis. The top set of rows of the table use the first error-prone proxy at both stages, the second set of rows use the mean of proxies at both stages, the third set of rows use the mean at the first stage and the first error-prone proxy at the second, and the final set of rows use the first error-prone proxy at the first stage and the mean at the second. Bold values indicate parameters for which the 95% percentile-based interval across the 1000 simulation replicates did **not** cover the true parameter value. The error models are specified to be normal (N), approximately normal (A), gamma (G), or uniform (U).

	Regression Calibration				Naive			
	A_1	A_1X_1	A_2	A_2X_2	A_1	A_1X_1	A_2	A_2X_2
N/N	1.0106	0.9961	1.0003	0.9995	1.0108	**0.8859**	1.045	**0.9101**
N/A	1.0175	0.9954	1.0007	0.9975	1.0174	**0.8248**	**1.0705**	**0.8586**
N/G	1.0109	1.0292	1.0036	1.0078	1.0103	**0.8577**	**1.0749**	**0.8647**
N/U	1.0062	1.034	1.0028	1.0226	1.0062	0.9729	1.0339	0.9616
A/A	1.0428	0.996	0.9989	0.9916	1.0414	**0.6132**	**1.1612**	**0.6666**
A/G	1.0187	**1.1453**	1.0069	1.0592	1.019	**0.7388**	**1.1893**	**0.7095**
A/U	1.0109	1.0405	1.0028	1.0318	1.0107	0.9659	1.0443	0.9495
G/G	**1.0866**	**1.2838**	**0.919**	**1.202**	**1.086**	**0.857**	**1.1308**	**0.8097**
G/U	1.0157	1.0608	0.9766	1.0494	1.0156	0.9842	1.0198	0.9638
U/U	1.0075	1.0367	0.991	1.043	1.0076	0.9957	1.0152	0.9957
N/N	1.0111	0.997	1.0013	0.9967	1.0111	**0.8872**	1.0454	**0.9074**
N/A	1.0173	0.9968	1.0002	0.9994	1.0165	**0.8249**	**1.0706**	**0.8586**
N/G	1.0247	1.0323	0.9832	1.0201	1.0248	**0.8602**	**1.0552**	**0.877**
N/U	1.0103	1.0358	0.995	1.0339	1.0098	0.9747	1.0269	0.9718
A/A	1.0393	1.008	0.9994	0.9984	1.0384	**0.6209**	**1.163**	**0.6743**
A/G	1.0537	**1.1959**	0.9615	**1.1174**	1.0534	**0.7688**	**1.1499**	**0.7544**
A/U	1.0104	1.0498	0.9957	1.0521	1.0101	0.9743	1.0383	0.9692

	Regression Calibration				Naive			
	A_1	A_1X_1	A_2	A_2X_2	A_1	A_1X_1	A_2	A_2X_2
G/G	**1.0951**	**1.4009**	**0.9042**	**1.3366**	**1.0947**	0.9351	**1.1268**	0.952
G/U	1.0161	1.0605	0.9816	1.0662	1.0159	0.9859	1.025	0.9803
U/U	1.0079	1.033	0.9914	1.0418	1.0077	0.9923	1.0154	0.9942
N/N	1.0116	0.9968	1.0005	0.9982	1.0118	**0.8866**	1.0453	**0.9097**
N/A	1.0177	0.9966	1.0008	0.9977	1.0171	**0.8257**	**1.0708**	**0.8578**
N/G	1.0263	1.0326	1.0028	1.0074	1.0262	**0.86**	**1.0744**	**0.8649**
N/U	1.0112	1.036	1.0029	1.0235	1.0112	0.9746	1.0344	0.9619
A/A	1.0372	1.0098	0.9988	0.9909	1.037	**0.6207**	**1.162**	**0.6673**
A/G	1.0495	**1.1943**	1.007	1.058	1.0493	**0.7689**	**1.1893**	**0.7076**
A/U	1.0116	1.0498	1.0027	1.0313	1.0117	0.9742	1.0453	0.9487
G/G	**1.072**	**1.4024**	**0.9193**	**1.2012**	**1.0719**	0.9346	**1.1294**	**0.809**
G/U	1.0161	1.0601	0.9755	1.0494	1.016	0.9857	1.0197	0.9645
U/U	1.0078	1.0329	0.9916	1.0433	1.0078	0.9924	1.0156	0.9957
N/N	1.0106	0.9963	1.0013	0.9974	1.0104	**0.8859**	1.0459	**0.9085**
N/A	1.017	0.9955	1.0009	0.9985	1.0166	**0.8249**	**1.0704**	**0.8581**
N/G	1.0095	1.0288	0.9834	1.0218	1.0089	**0.8568**	**1.0569**	**0.8763**
N/U	1.0047	1.034	0.9952	1.0332	1.0048	0.9736	1.0266	0.9711
A/A	1.0441	0.997	0.9987	0.9996	1.0439	**0.6129**	**1.1602**	**0.6746**
A/G	1.0232	**1.1462**	0.9612	**1.1165**	1.023	**0.7378**	**1.1498**	**0.7538**
A/U	1.0101	1.0408	0.9958	1.0529	1.0102	0.9655	1.0377	0.969
G/G	**1.1095**	**1.2825**	**0.904**	**1.337**	**1.1087**	**0.8556**	**1.1267**	0.9509
G/U	1.016	1.0603	0.9816	1.0659	1.0159	0.9845	1.0256	0.9818
U/U	1.0074	1.0369	0.9916	1.0424	1.0074	0.9957	1.0157	0.9946